CHAPTER II

REVIEW OF THE LITERATURE

2.1 Introduction

This study aims to investigate the level of Thai graduating students' listening ability in English for the service and hospitality industry. It will further study the relationship between the TOEIC Test and the new test (L-PESH Test) to see whether the L-PESH Test can assess the students' listening ability.

The findings in this study can help identify the level of listening ability of graduating students before entering the job market. The universities can use the test as an indicator of how to improve the English courses and how to increase the level of their students' ability in listening. The test can also be applied as a tool in improving one's listening ability in the language self-study centre. Moreover, the employers can use the test results as a qualification in selecting employees or arranging staff training.

This chapter sets out a review of related literature beginning with characteristics of hospitality language followed by the significance of listening ability in communication and listening proficiency. Next, it discusses certain theoretical aspects of language testing including language proficiency testing, an overview of listening tests, testing listening skills, English for Specific Purposes and its testing, a framework for analyzing Target Language Use (TLU) and test task characteristics, test specifications, and analysis of test task characteristics. Finally, the chapter presents recent studies on how cut-off scores and their ability descriptors can be established and validated.

2.2 Characteristics of hospitality language

As the tourism and hotel industry has developed; a differentiation in hosting activities has arisen, between those that are extended as a social obligation and those involving payment. In both categories, participants normally observe the etiquette

and proprieties that are traditionally practiced, and both involve interpersonal and, in some case, cross-cultural communication (Blue and Harun, 2003: 74).

As the concern in this study is primarily with "commercial hospitality", hospitality refers to the cluster of activities oriented towards satisfying guests. To hoteliers, hospitality means looking after the guests well; therefore, hospitality language refers to all linguistic expressions which relate to and represent hospitality concerns. The language is formal, though it depends very much on the level of acquaintance among participants. Hospitality language has a long history of development. Thus, there is a wide variety of terms dealing with hospitality in many cultures; the hospitality register for the English language is quite extensive (Blue and Harun, 2003:74).

In addition, English is the most commonly used language of hospitality and the lingua franca of tourists and travellers worldwide. Therefore, in many parts of the world, the art of greeting, soliciting information, thanking and bidding farewell requires some measures of familiarization with the relevant English expressions before a person can serve effectively as a receptionist, telephone operator, or in other guest-contact capacities.

Thus, whether in English or another language, there is an identifiable cluster of language skills which staff dealing with hotel guests should have already acquired. At the very minimum, these skills include:

- how to address a person;
- how to solicit and give necessary information;
- how to respond to questions/requests;
- how to use prompts;
- how to use gestures;
- how to deal with difficult customers;
- how to deal with complaints.

Blue and Harun (2003) further explain that hospitality language, viewed as a process, covers at least four stages: arrival, familiarization, engagement and

departure. Each different situation warrants different types of hospitality, and the cycle does not always follow the same sequence.

Kasavana (1993: 424 cited in Blue and Harun, 2003: 75) has illustrated the cycle of hospitality practices in hotels in Table 2.1 This cycle is also known as "guest cycle".

Table 2.1: The Commercial Arrival-Departure Hospitality Cycle

Stage	Activity	Language used
Arrival	- Picking-up service in some hotels. - Luggage transfer. - At the reception. - All services in this stage are commercial.	- Routine and rehearsed language used. - Formal question-answer transactions in a formal tone. - Varies with category of hotel. The examples include greeting by driver and welcoming by receptionists.
Familiarization	- Briefing the guests on what and where in-house facilities are available, and on meal and checkout times. - Guests' Reading in-house brochures and asking questions about hotel.	- Briefing style of language. - Rehearsed messages. - Additional questions and answers. - Formal tone language use varies according to category of hotel.
Engagement	- Independent use of facilities in rooms and in different sections of the hotel.	- Mostly formal and impersonal, but may depend on how long guest stays in a hotel. Difficult to predict exact language needs other than those relating to use of facilities.
Departure	- Luggage transfer. - Preparation of bill. - Farewell conversation.	- Mostly rehearsed language, - - Mostly formal and impersonal language.

From the previous review, it is obvious that anyone who wishes to get a job in this industry needs to have a certain level of English proficiency, especially in listening skills.

Making people feel welcome is indeed an art, and a key to success in the service and hospitality industry. It has now become a standard feature of commercial hospitality practices (Blue and Harun, 2003). In the context of a globalized world, there has been some standardization of hospitality language. The language of hotel encounters comprises functional aspects of hospitality language that are understood worldwide. These functional activities include check-ins, checkouts, information and queries, and miscellaneous requests.

2.3 Significance of listening ability in communication

Listening skill is crucial in a career of service and hospitality. The hotel personnel have to be good at listening in English. Many scholars and educators have elaborated the significance of listening proficiency. Hunt (1987:14) states that in 70% of working hours, people listen more than speak in order to communicate. Lundsteen (1990:213 cited in Kreutanu, 1998:21) adds that in everyday communication listening skill is mostly used. Listening is a key in learning and initiate interactions among people. Oxford (1993:206) points out that among the four skills in English, listening is more important and plays more significant roles in communication than the other three skills. That is, people spend 45% of their time on listening, 30% in speaking, 16% in reading and 9% in writing.

In business, listening proficiency even plays more significant roles, as stated in Abrams (1986:77) because listening is essential in doing business. Business people manage their business through listening as in telephoning, negotiating, making appointments, ordering goods, and so on. If these people do not have good listening ability, the business may not run as smoothly as in those who have higher proficiency in English listening. For example, if hotel reservation staffs do not have good listening skills, they can make mistakes about time and dates in taking reservations. The customers will be dissatisfied and the hotel can lose a lot of benefits from this.

Moreover, listening ability is also important in learning English as a second and a foreign language. This is supported by a summary of significance of listening ability on learning English as a second or foreign language by Rost, 1994 (quoted in Nunan and Miller, 1995:5)

- Listening is essential in language learning as it acts as language input to the learners. If the learners cannot comprehend what has been input, they can learn nothing.
- Listening starts interaction. When the learners understand what is heard and can interact, then language learning occurs.
- It is challenging for language learners to comprehend what is heard from authentic listening situations.
- Listening practices include many interesting activities that motivate learners to learn more.

To sum up, listening plays a crucial role in communication. Those who plan to get a job in the area of service and hospitality need to have adequate listening proficiency in order to perform their jobs effectively.

2.4 Listening proficiency

Kreutanu (1998:25) explains that listening proficiency is the ability to understand words and information heard and to summarize the main idea of what is heard using both language knowledge and background knowledge in order to comprehend what is heard. Brown and Yule (1988:58) add that the listeners need knowledge of the language and context to interpret what is heard to link it together so that the information heard can be comprehended.

McKeating (1985:59-63) further indicates that perception and decoding are two factors that build up listening proficiency in English. To understand short listening information, first the listeners perceive different sounds, words, and phrases and then understand them. Next, the listeners store information that has been perceived in short-term memory as much as they can. Finally, they relate what is heard to their background knowledge and context in order to understand it.

In order to enable students to communicate effectively in English, the students should be trained in all English skills; listening, speaking, reading, and writing. Among these four skills, listening is the most important skill in communication as it is mostly used in daily communication. Besides, listening is the foundation skill to further learning in the other three skills. (Pongkasempornkul, 1998:4).

Kanchanasathit (1980: 276-281) found from her survey on the needs of English used in business fields that business people need specific training in business English, especially training that improve their listening and speaking skills.

Based on the literature reviewed in the previous section, the following conclusions are drawn. First, listening skill is essential in both language learning and business. Second, listening is the skill that needs training and improvements most. Finally, listening is the skill used most in learning and communication. Therefore, in this study the researcher will focus attention only on listening skills.

2.5 Language proficiency tests

Hughes (1989) explained that proficiency tests are designed to measure people's ability in a language regardless of any training they may have had in that language. Therefore, the content of the test is not based on the course content, but rather on what the candidates can perform in the language in order to be considered proficient. According to Hughes (1989:9), proficient means "*having sufficient command of the language for a particular purpose*".

Heaton (1990: 17) adds that we use proficiency tests to measure suitable candidates in performing a certain task. The test can measure candidates' proficiency in certain special fields such as medicine, social studies, physical sciences, technology, and so on. As a result, some proficiency tests concentrate on assessing the candidate's ability to use English for specific purposes.

In this way, a proficiency test is able to measure the actual ways in which English will be used in the future. When designing a proficiency test, we should pay careful attention to the language areas and skills that the candidates will need.

2.6 An overview of listening tests

Listening comprehension is very difficult to describe and to assess because it is an invisible cognitive operation. Despite the lack of adequate theories and models of listening, numerous attempts have been made to describe listening ability (Brindley, 1997: 65).

Buck (2001: 1) adds that listening comprehension is a very complex process. To measure this process, first, the test developer has to understand how the process works. An understanding of what to measure, a construct, is the starting point for test construction. Therefore, the first task of the test developer is to understand the test construct and then to make a test that somehow measures that construct. This can be referred to as construct validity. The central issue in all assessment is ensuring that the right construct is being measured.

2.6.1 Different types of knowledge used in listening

Buck (2001:1) explains that to consider how the language comprehension system works, it is obvious that a number of different types of knowledge, both linguistic and non-linguistic knowledge, are involved. The most important linguistic knowledge is phonology, lexis, syntax, semantics and discourse structure. The non-linguistic knowledge is the knowledge about the topic, about the context, and general knowledge about the world and how it works.

2.6.2 Processes in second language listening:

According to Brindley (1997), listening processes in second language are categorized into two main groups as presented in the following paragraphs.

2.6.2.1 The bottom-up process

In the bottom-up process, the smallest units of language are identified first, and then are chained together to form the next highest unit, then again are put together to form the next highest unit (Brindley, 1997:2).

When people start thinking about language processing, they often assume that the process occurs in a definite order, starting with the lowest level of detail and moving up to the highest level. Therefore, they assume that the listening input is first decoded into phonemes (the smallest sound segments that can carry meaning), and then this knowledge is used to identify words, then the processing continues to the next higher stage, an analysis of the semantic content and understanding of the basic linguistics meaning, and finally understand what the speakers mean.

However, Buck (2001:2) argues that there are some serious problems with this process of language comprehension as both research and daily experience indicate that the processing of different types of knowledge does not occur in a fixed sequence, but rather that different types of processing may occur simultaneously.

2.6.2.2 The top-down process

On the other hand, the top-down process depends on the use of context and background knowledge to understand the meaning of the incoming message (Brindley, 1997: 67). Buck (2001:3) believes that listening comprehension is a top-down process in the sense that various types of knowledge involved in understanding a language are not applied in any fixed order. He says that this can be referred to as an interactive process.

Brindley (1997: 67) points out that in recent years, simple bottom-up or top-down processes have been rejected as inadequate for explaining how second language learners acquire input. Interactive-compensatory models that are based on the view that information from more than one level is utilized simultaneously have replaced them.

2.7 Factors affecting second language listening

Brindley (1997) gathered crucial factors, suggested by various authors that affect task difficulty of listening tests as follows:

Lexical knowledge - Lack of knowledge of key lexis can lead to miscommunication or even breakdown (Brindley, 1997:68). The ignorance of vocabulary was the major factor that causes lack of listening comprehension beyond the intermediate level of language learning.

Syntactic knowledge - Target language syntax seems to be an important factor in increasing the amount of linguistic material that can be retained in short-term memory.

Background knowledge - It is suggested that learners' background knowledge is of major importance in determining how the message heard can be interpreted. Background knowledge is obviously important in listening comprehension. If the listener shares the same knowledge as the speaker, much of what is said can be understood in terms of the top-down process.

Speech rate - A range of studies on the speech rate in SLA listening tests has been reviewed and it was found that the faster rates of delivery can significantly reduce comprehension.

Noise - The ability to understand the message of native and non-native learners when the noise ratio increases is different.

Contextual support- It is a very important fact affecting second language listening, particularly at the lower levels of ability.

Memory- It is obviously an important factor in language comprehension.

2.8 Approaches to assessing listening

The basic task in making assessment is to take theoretical notions about a construct and operate them, that is, to turn them into actual practice, in the form of test items. Historically, there have been three main approaches to language testing: the discrete-point, integrative, and communicative approaches. Theoretical notions underlying testing practice and certain testing techniques are associated with these approaches.

Buck (2001) has summarized those three approaches and their testing techniques as follows:

2.8.1 The discrete-point approach

As stated in Buck (2001), the most famous advocate of this approach was Lado, who defined language in behaviorist terms. He considered listening comprehension to be a process of recognizing the sounds of a language. Thus, testing listening comprehension means testing the ability to recognize elements of the language in their oral form. Generally, discrete-point tests use selected responses such as true/false and the three- or four- option multiple choices respectively. Discrete-point items are usually scored by giving one point for each correct item. The most common tasks for testing listening in this approach are:

- **Phonemes discrimination task**, often the words are in minimal pairs

  Example: Test takers hear:

  " *I hear they have developed a better vine near here.* "

  They read:

  I hear they have developed a better vine/wine near here.

  (Buck, 2001:63)

  In this example the test takers listen and choose the correct words they hear.

- **Paraphrase recognition**

Example: Test takers hear:

" John ran into a classmate on his way to the library. "

They read:

a. John exercised with his classmate.

b. John ran to the library.

c. John injured his classmate with his car.

d. John unexpectedly met a classmate.

(Buck, 2001:63)

In this example the test takers listen to a statement and then choose the option that is closest in meaning.

- **Response evaluation**, which is similar to what Heaton (1976:76) called "statements and dialogues". This technique tests the ability to understand both the grammatical and lexical features of a short utterance.

Example Test takers hear:

Man: *"You're never ready on time, Mary. I'm sure we've missed the bus to town."*

Woman: *" Let's not argue here, Bill. Look. That's Mr. Green's car. He's stopping to give us a lift. And there's the bus. We can take our pick now!"*

They read:

a. Mary and Bill have missed the bus to town.

b. Mary and Bill are arguing about Mr. Green's car.

c. Bill is always late because he likes to pick flowers for Mary.

d. Mary and Bill can go to town either by bus or in Mr. Green's car.

(Heaton, 1976:77-78)

In this example, the test takers hear statements or a brief dialogue and a question concerning the dialogue then choose the best response. The question may

test straightforward comprehension of the dialogue or test the test taker's ability to make deductions or draw inferences from the dialogue.

2.8.2 Integrative testing

Oller (1979 cited in Buck 2001) has proposed the idea of "integrative tests". He explains that

> *"whereas discrete items attempt to test knowledge of language one bit at a time , integrative tests attempt to assess a learner's capacity to use many bits all at the same time"*
>
> (Oller 1979:37 cited in Buck 2001)

Oller has based his ideas on what he called a "pragmatic expectancy grammar". This means that there are regular and rule-governed relationships between the various elements of the language, and to know a language it is necessary to know how these elements relate to each other. Redundancy is also an important way the elements of language relate to each other. It is useful to see exactly how redundancy works.

He further explains that because of redundancy, those who know the language well will be able to make predictions about the language based on this pragmatic expectancy grammar and the ability to make predictions can be used to measure proficiency in the language. Many tests of reduced redundancy, in which elements are removed thus reducing the redundancy of the text, have become widely used and are closely related to integrative testing. The following are examples of integrative tests.

- **Noise tests**: In this test the test takers listen to a passage which has been mutilated by addition of background noise, white noise, meaning that the noise which covers most of the frequency range of spectrum, a sort of continuous hiss. These white noises mask the text. The test-takers listen and respond by repeating what they have heard either speaking aloud, or writing it down during appropriate pauses.

- **Listening cloze**: The cloze tests are reading tests based on the idea of reduced redundancy. This technique is now widely adapted to listening tests. Templeton (1977 cited in Buck 2001) has reported the result of his study that listening cloze has high validity on both theoretical and practical grounds. He suggests that the test takes little administration time and item writing is easy. The only problem found is the technical problem of making the recording using traditional analogue tape. However, this problem can be solved by using digital audio-editing software that is now cheap and easy to use. Later on, this technique is modified to a gap-filling technique. However, there is a question whether the test-takers can fill in the blanks based on their comprehension or guessing without listening to the passage. To prevent this problem and force test-takers to process the meaning in order to fill in the blanks, gap-filling summaries are introduced.

- **Gap-filling summaries:** In this technique, the test-takers are given a summary of the passage they are about to hear, in which some important content words have been replaced by blanks. After looking at the summary for a while, test takers then listen to the original passage. Their task is to use their understanding of the passage to fill in the blanks.

Buck (2001) further explains that this technique can be used as an integrative test of short sections of text, but the gaps can also be selected so that the test- takers are required to understand discourse features, summarize parts of the text or make inference about the overall meaning.

- **Dictation:** This technique is the most widely used integrative test of listening. The test-takers listen to a passage and write down what they have heard. Usually they listen to the passage twice: in the first time just listen and try to understand the passage. The second time, the passage is broken into a number of short segments, with a pause between each. During the pause, the test-takers write down what they have heard. When scoring dictation, remember that they are not designed to be the test of

spelling. Spelling mistakes should therefore be ignored. Thus, a better way to score dictation is to delete marks.

However, Hughes (1989:139) found that scoring dictation for low-ability test–takers can be very difficult when they make a lot of mistakes. He then recommends "partial dictation", in which some of the text is already written down on the answer sheet. This technique makes it easier for them to keep oriented within the passage.

Example -First, test-takers hear the original passage.
-Then, they listen to the text again and write the underlined part; the other part is already provided in written English.

Segment:	#words
I am an English man and I live in Japan. One thing	
which I think is very strange	*[6]*
is the use of microphones on Japanese television.	
On English television you never see a microphone.	*[8]*
They are always hidden.	

(Buck, 2001:76)

- **Sentence-repetition tasks:** This technique is basically the same as a dictation, except that test-takers repeat the text orally during the pause between each section. Usually, they are given a series of unconnected sentences rather than a unified passage. They hear each sentence once and repeat it back immediately after they have heard it. The responses are often tape-recorded and then scored later. Buck (2001) argues that sentence-repetition tasks are not just tests of listening, but also tests for general oral skills.

- **Translation**: This technique is not usually considered an integrative test. In this technique, the test-takers are asked to listen to the recording and, then during the pauses, write down a translation of what the passage has stated.

2.9 Testing Listening Skills

Before we look at ways of testing listening skills, it is necessary that we understand the characteristics of the spoken language and its implication for listening tests.

2.9.1 Characteristics of the spoken language and its implication for listening tests

According to Heaton (1990:41-45), spoken language contains a lot of redundancy, and meaning is usually reinforced and repeated in several ways. The message can still be comprehended even when several words are omitted. When speaking, the speakers often hesitate and pause, filling in the gaps with sounds like "er", em", and so on. Sometimes they have false starts, start a sentence, change their mind and start it again.

The structure of spoken and written language is different. In writing, the writers organize language in sentences, but in spoken, speakers organize language in clauses, and seldom use complex sentences in spontaneous speech.

Moreover, the language is usually presented in a certain situation for a particular purpose and people remember the general meaning of a sentence rather than the actual word. Thus, it is important that the teachers educate students that it is not essential to understand every word in order to understand the overall meaning.

In addition, the use of gestures, eye contact, and facial expressions also help listeners to understand a message better. Thus it is always more difficult to understand someone talking over the telephone, or conversations recorded on cassettes. Therefore, a higher and more intense level of listening ability is often demanded.

2.9.2 Implications for listening tests

In a test of listening, it is better to talk rather than to read aloud long written texts to the students as the written texts lack most of the redundant features that are

so important in helping the students understand the speech. However, in practice, it is usually difficult to do so in a foreign language. Some teachers may not feel confident enough to give spontaneous talk. Consequently, they read aloud a written text. To make reading aloud more like spoken language the teachers can rewrite most of the complex sentences into fairly short phrases or clauses, using coordinating conjunctions as and, but, or, so instead of subordinating ones like although, whereas, in order to, and so forth. The texts should be read at natural speed with slightly longer pauses at the end of clauses and sentences.

In most tests, teachers use recordings in the test. Talks and conversations recorded on cassette tapes are even harder tests of understanding than those given in real life. However, cassette tape recordings have the following advantages.
First, they make a listening test more reliable as the same voice is heard in exactly the same way regardless of number of times the test is given. Next, it is possible to use the recorded voices of native-speakers or other non-native speakers of English in the recordings. Finally, it is possible to play recordings of conversations involving two or more speakers at a time, instead of having one teacher read aloud the voices of different speakers.

2.9.3 Short statements and conversations

As listening to long talks in a foreign language demands higher and intense ability, it is generally more appropriate to let learners at early stages listen to short statements and conversations. Therefore, a number of listening tests contain short statements and conversations in the form of instructions or directions, short conversations on which questions or tables, or pictures are based.

2.9.4 Longer conversations and talks

When giving talks, short lectures, or longer conversations for listening comprehension, remember that the test should not become a test of memory (Heaton, 1990: 53). It is also suggested that in some cases the students are given incomplete notes and finish the notes as they listen to the talk. In addition, completing tables, writing true/false about the conversations are not too difficult and suggested for the

listening comprehension tests. It is also useful to vary the materials in a listening comprehension test including both talks and conversations.

2.9.5 Four categories of testing listening skills

According to Heaton (1990), testing listening skills can be grouped into four categories:

2.9.5.1 Distinguishing between sounds: At early stages of learning English students will probably have some difficulty in hearing the difference between one particular sound and another. For example, they may not be able to tell the difference between "live" and "leave" or "raw" and "law". In teaching and testing the ability to recognize the different sounds, it is easier to start by pronouncing the words in isolation. Then say the words naturally in sentences and test sound differences in context.

Tests of stress and intonation are very important at the early level too. Though some students have learned the correct stress patterns for certain words, they still cannot pronounce them correctly. Thus, it seems of little use learning to recognize word stress without being able to apply this knowledge (Heaton, 1990). Moreover, testing stress and intonation is often artificial. It is far better to concentrate on testing students' understanding of short conversations and talks in as natural ways as possible. It is also important to realize that the ability to hear sound differences is not as necessary as the ability to understand spoken messages.

2.9.5.2 Dictation: Some teachers have considered dictation mainly as a test of spelling; in fact, it tests a wide range of skills. It can provide a useful means of measuring general language performance. Dictation has long been closely associated with listening comprehension.

When you give a dictation to your class, begin by reading through the whole dictation passage at normal speed. Then dictate meaningful units of words (phrases and short clauses) reading them aloud as clearly as possible. Finally, after finishing the actual dictation of various phrases and clauses, read the whole passage once more

at slightly slower than normal speed. Give appropriate time for students to check the spelling of words and their overall understanding of the text.

However, in practice, many teachers try to make the dictation easier for their students by reading the text very slowly, word by word. This can be harmful as it encourages students to concentrate on single words. It is also necessary that the text for dictation be prepared beforehand.

Marking dictation is fairly straightforward, usually done by deducting half a mark or one mark for each error. It is also useful if the teachers can use the same text first for listening comprehension and then for dictation.

2.9.5.3 Testing listening comprehension

As suggested in Heaton (1990), listening comprehension can be assessed in many ways. For example, listening comprehension can be assessed through visual materials, through statements and dialogues, and through talks and lectures.

2.9.5.4 Testing listening with other skills

There are several other types of multi-mode tests (Heaton, 1990:54) involving listening. The most frequently found is the combination of a speaking test with listening for example, when having a conversation, it is necessary to listen before speaking. Listening is an integral part of speaking in everyday life. This kind of test is often referred to as "oral interview" or "listening-speaking test". Listening can also be combined with writing in several ways such as listen and take notes, listen and give short answers, listen and fill in the gaps, and so forth. Though listening can be integrated with other skills tests, Heaton (1990) noted that listening skills could be best developed, taught, and tested on their own as skills not dependent on other language skills.

2.10 English for Specific Purposes and its test

Why test language for specific purposes?

According to Douglas (2000), there are two main reasons why we do not use an existing, general-purpose language tests such as TOEFL, IELTS, CPE, etc. for specific test- takers.

Reason One: Language performances vary with context.

Researchers are pretty much in agreement that language performances vary with both context and test task, and therefore our interpretations of a test taker's language ability must vary from performance to performance. However, it is not enough merely to give test takers topics relevant to the field they are studying or working in: the material the test is based on must be authentic and provide a task in which both language ability and knowledge of the field interact in a way which is similar to the target language use situation.

Reason Two: Specific purpose language is precise.

A second reason for preferring Language for Specific Purposes (LSP) tests to more general is that technical language has specific characteristics, and specific communicative functions within a field, namely "precision", that people who work in the field must control. And it is this precision that is a major focus of specific purpose language use.

Douglas (2000:19) has proposed a more precise definition of specific language tests as follows:

A specific purpose language test is one in which test content and methods are derived from an analysis of a specific purpose target language use situation, so that test tasks and content are authentically representative to tasks in the target situation, allowing for an interaction between the test taker's language ability and specific purpose content knowledge, on the one hand, and the test tasks on the other. Such a test allows us to make inferences about a test takers' capacity to use language in the specific purpose domain (Douglas, 2000: 19).

2.11 A framework for analysing TLU and test task characteristics

Douglas (2000) has used the term TLU (Target Language Use) to echo the term "target language use domain" used by Bachman and Palmer(1996) who define it as

> *"a set of specific language use tasks that the test taker is likely to encounter outside of the test itself, and to which we want our inferences about language ability to generalize."*
>
> (Bachman and Palmer, 1996:44)

This framework allows test developers to analyze a TLU situation and to develop test tasks that reflect the characteristics of the target situation in order to provide a basis for test development (Douglas, 2000:50). These characteristics include the features of rubric, the input, the expected response to the input, the interaction between input and response, and assessment criteria.

Moreover, this framework helps to describe the features of tasks in both the specific language use situation in which language testers are interested and the language test they wish to develop.

Language test developers need this framework because they have to be certain that the specific purpose context and the test share essential characteristics so that the test taker's performances on the test tasks can be interpreted as evidence of their ability to perform tasks in non-test TLU situations.

Bachman (1990:112) mentioned that the correspondence between the TLU contexts and the methods used to measure language ability would affect the authenticity of test performances. Moreover, the closer the correspondence between the TLU contexts, the more authentic the test task will be for the test takers.

The framework used to describe the TLU context and TLU test task characteristics in Listening Proficiency Test in English for Service and Hospitality

Industry (L-PESH Test) is based on Bachman and Palmer's (1996) together with some adaptations by Douglas (2000) to fit a particular case of LSP testing.

Douglas has also noted that among those characteristics, the input and the expected response are more relevant to the LSP testing than others because they can prevent any test takers from performing below their ability. The test takers may not be adequately familiar with the procedures for responding to the tasks or they may misunderstand the criteria on which their response would be judged. Therefore, care must be taken that all of test takers follow the same procedures for responding to the test tasks. (Douglas, 2000: 49)

However, it is true in principle that all characteristics of test tasks are derivable from the specific purpose TLU situation. The degree to which this is true in practice will vary with external considerations often unrelated to the target situation.

2.11.1 The characteristics of the rubric:

In testing, the term "rubric" has been defined by Bachman (1990:118) as "characteristics specify how test takers are expected to proceed in taking the test; and include the instructions, time allocation, and test organization. Douglas (2000) has based his framework on Bachman and Palmer's (1996) with slight changes in order to relate to the nature of the communicative event in the TLU situation and in language for specific purposes tests. Therefore, the characteristics of the rubric in this framework include:

-The specifications of objectives or the purposes of the test,

-The procedures for responding including information about how the test takers are to respond to the test tasks. In non-test situations, this information will usually be implicit in the situation, whereas in test situations, this information must be given explicitly. For example, the test takers are told to respond by checking boxes, writing words in the blanks, filling out a table or form, and so forth.

- The structure of the communicative event including number of tasks, relative importance of tasks, and distinction between tasks.

-The time allotment stating how much time the test takers have to perform the test tasks. This information is usually given explicitly in both test and non-test situations. For example, in a test situation, the test taker is told that "You will have 20 minutes to complete this section of the test", and in a non-test situation, an employee is told that "Anna, please have the minutes of the meeting on my desk by 4 pm."

-The characteristics of evaluation including criteria for correctness and procedures for rating the performance. The test takers are explicitly told about criteria by which their language performance will be judged.

2.11.2 The characteristics of the input:

These characteristics refer to the specific information that test takers must process in dealing with a communicative situation. Two types of data: prompt data and input data are included in these characteristics. The prompt data provides information about features of the LSP context consisting of the setting, participants, purpose, form and content, tone, language, norms of interaction, genre, and problems. The input data provides information about the format of the test and its level of authenticity. The format data will be analyzed in terms of whether the test is visual or auditory (or both), the means by which it is delivered, and its length in terms of time and number of words. The level of authenticity refers to both the degree to which the input data in a test reflects the characteristics of the TLU situation, and the degree to which the data engages the test taker's communicative language ability. (Douglas, 2000:73).

The prompt and input data are clearly very important aspects of an LSP test in establishing the specific purpose context for the test taker. The information provided in the prompt and the features data need to be as rich and engaging as possible (Douglas, 2000: 59).

2.11.3 The characteristics of the expected response:

They are those relevant to the format of the response that can be written, spoken, physical, or a combination of these. The type of response can be a selected response, a limited production response, or an extended production response. Response content may include the features of the nature of the language and background knowledge to be assessed, and the level of authenticity-both situational and interactional.

2.11.4 Characteristics of the interaction between input and response:

Bachman and Palmer (1996) pointed out that in normal language use situations, input and response interact along at least three dimensions: reactivity, scope, and directness. Douglas (2000) further explained these dimensions as applied to LSP testing.

2.11.5 Characteristics of the assessment:

These include (1) the construct definition, "a statement of what aspects of specific purpose language ability are to be measured, usually derived from an analysis of the TLU situation" (Douglas, 2000:74), (2) the criteria for correctness, and (3) a systematic set of procedures for carrying out the rating/scoring of the performance. LSP test developers must ensure that the construct definition and assessment criteria are clearly, completely, and precisely stated.

2.12 Techniques for investigating the target language use domain

The general problem in LSP testing is usually the case that the testers are seldom experts in the field in which they are attempting to measure language ability and many seek expert's help to understand the TLU situation and the characteristics of input data to be used as the basis of the LSP test. Douglas (2000) thus suggests two approaches to investigate and describe the TLU situations to be translated into test tasks as follows:

2.12.1 Grounded ethnography and context-based research – This technique involves obtaining commentary from participants in the language use situation on records such as videotapes, audiotapes, and written documents. The information obtained is used as the basis for understanding what in the performance is interesting, noteworthy, or problematic from the points of view of participants. It is also important that field specific test developers understand the TLU situation from the perspective of language users in that domain.

2.12.2 Subject specialist informant procedures – Despite the fact that the LSP test developers are able to obtain satisfactory recordings of information on specific purpose TLU situations and samples of authentic input data, there still remains the difficulty of determining what in the data is worth focusing on. This inability to determine the focus may also be caused partly by the fact that the test developers are usually someone who knows little about the specific purpose situations in which they are working. It is thus essential that the LSP testers make use of specialists informants in a principled way in analysing the TLU situation during the test development process.

2.13 Test Specifications

According to Douglas (2000:109) the term "test specification" usually refers to a "document that serves as a kind of blueprint for test developers and item writers, a reference point for validation researchers, and sometimes a source of information for score users". The test specification document is an indispensable part of the test development process. It serves as a guide for the construction of the test.

> *An essential and the most difficult aspect of producing test specifications is making a leap from the analysis of the TLU tasks to the specification of the test tasks. Translating the TLU task characteristics into LSP test tasks requires a large amount of judgment and the weighing of alternatives, often making compromises based on practical considerations related to budgetary and time constraints.*
>
> (Douglas, 2000:113)

Bachman and Palmer (1996: 106) also pointed out that "not all tasks will be appropriate for use as a basis for the development of test tasks" because they may not meet all the criteria for good testing practice. "Some tasks may be so highly fielded specific that raters cannot be trained to reliably assess language ability in such context-embedded performances. In the case of validity considerations, some TLU tasks may be inappropriate sources of information for types of inferences we wish to make about the test takers' language ability. In terms of practicality, a TLU task may require more equipment, personnel or time than is reasonably available in the test situation.

Thus, in making the transition from analysis of the target language use tasks to test tasks, we must bear in mind the qualities of test usefulness and often we are required to adapt TLU tasks to the test situation.

2.14 Essential components of LSP specifications

The following are components of test specifications proposed by Douglas (2000):

2.14.1 The purposes of the test: including explicit decisions we want to make based upon inferences about language ability or capacity for language use and outline any constraints on the test situation.

2.14.2 The TLU situation and list of the TLU tasks: involving the domain in which inferences about language ability or capacity for language use will be made.

2.14.3 The characteristics of the language users / test takers: describing the nature of the population for which the test is being designed.

2.14.4 The construct to be measured: describing the nature of the ability to be measured, providing a description of precise specifications of specific purpose language ability for making inferences about the result of test performance.

2.14.5 The content of the test: specifying types of test tasks, organization of the test, number of tasks, brief description of each task, time

allocation for tasks and the length of text included in the task, and specifications of test task.

2.14.6 Criteria for correctness: providing description of how responses will be judged correct, how to assign levels on a rating scale, and how a total score will be calculated.

2.14.7 Samples of tasks / items: demonstrating test items or test tasks.

2.14.8 Plan or evaluating the qualities of good testing practice: including the following:

- Validity: the interpretations made on test performance,
- Reliability: the consistency and accuracy of the measurements,
- Situational authenticity: the relationship between the target situation and the test tasks,
- International authenticity: the engagement of the test taker's communicative language ability,
- Impact: the influence the test has on learners, teachers, and educational systems,
- Practicality: the constraints imposed by such factors as money, time, personnel, and educational policies.

In conclusion, to develop a language for specific purpose test, the test developers need to describe the features of tasks in specific language use situations. They have to be certain that the specific purpose context and the test share essential characteristics so that the test taker's performances on the test tasks can be interpreted as evidence of their ability to perform tasks in non-test TLU situations. Thus, in making the transition from analysis of the target language use tasks to test tasks, test developers must bear in mind the qualities of test usefulness and often adapts TLU tasks to the test situation. In addition, prior to the test- item writing, the test developers need to design the test specifications. The test specification document is an indispensable part of the test development process. It serves as a guide for the construction of the test.

2.15 Standard setting- identifying cut-off scores and their descriptors

One of the purposes of standardized tests is to increase accountability among educators, students, testers, test takers, and related stakeholders. As such, students and test takers are expected to meet a standard of proficiency that the tests are designed to assess. The standard should represent mastery of the learning objectives, or a level of basic proficiency necessary to move on to the next level, or to function in the real world (van der Linden, 1982).

Once established, a standard is translated into a cut-off score in the distribution of scores obtained from a set of test items relevant to and representative of the standard. The method used to set these cut-off scores is called standard setting. The purpose of the cut-off score is to separate test takers who meet the standard from those who do not (Morgan and Michaelides, 2005). However, once the need to establish a performance standard, or to set cut-off scores, has been established, the following question arises: What is the best method to use to set cut scores?

There are several acceptable methods to set cut-off scores; for example, the Angoff method, the Modified Angoff method, the Bookmark method, and so forth. Each method depends on having participants in the standard setting who are very knowledgeable about the test content standards and willing to help define the level of knowledge and skill expected of a test taker at each performance level articulated by the cut-off scores. Morgan and Michaelides (2005:1) further explain that no one standard-setting method is agreed upon as the best. Because it is possible that different standard-setting methods may result in different recommended cut-off scores, it is essential that careful thought goes into the decision of which standard-setting method to use. Part of this thought process should include consideration of the arguments defending the validity of the standard-setting method for the purpose of which the resulting cut-off scores will be used. Additional thought should be given to the type of evidence or documentation that should be collected and maintained during the process of setting cut-off scores.

Assessments may be composed of a variety of item types; for example, those scored dichotomously i.e., multiple-choice, true-false, and other items with clear right or wrong responses, and those scored polytomously i.e., essays, performance tasks, open-ended items or some short-response items where it is possible to receive partial credit for a correct but incomplete response (Morgan and Michaelides, 2005:1). A variety of standard-setting methods have been developed. However, many of the methods work best with a particular item type, and thus matching the test format to an appropriate method should help determine which standard-setting method will be used or, at the very least, which methods will not be used. For example,

> *"the Modified Angoff method (Angoff, 1971) has a long history of use in setting cut scores for tests with primarily multiple-choice or dichotomous items. Hambleton and Plake (1995) provided extensions to the Modified Angoff method for its application to performance-based tasks. The Body of Work method (Kahl, Crockett, DePascale, and Rindfleisch, 1994, 1995; Kingston, Kahl, Sweeney, and Bay, 2001) is a more recent method for setting cut scores but is designed for assessments with more open-ended tasks and fewer dichotomous items."*
>
> (Cited in Morgan and Michaelides (2005:1).

2.15.1 What are cut-off scores?

As defined by a number of relevant experts, a cut-off score represents a standard of performance that is set in a selection process with the objective of identifying the best-qualified candidates. In setting a cut-off score, the testers are deciding on the level of performance that a test taker must display to be considered further.

Biddle (1993) further explains that the specific score that is used as the cut-off score is what separates those who pass a test from those who do not. It is this score that determines the consequences of taking the test. For example, if the tester scores below the cut off it will mean having to try again for certification.

2.15.2 Types of cut-off scores

From related literature reviews, it can be concluded that setting cut-off scores may be divided into two major types: performance related and group related.

2.15.2.1 Performance-related cut-off scores

Performance-related cut-off scores are set by making a judgement about the test score or the level of the qualification that corresponds to the desired level of job performance.

2.15.2.2 Group-related cut-off scores

Group-related cut-off scores are set relative to the performance of the candidates in a reference group.

Both methods may be used in combination in order to select the highest ranking candidates while ensuring that they demonstrate a minimum level of performance on the test.

2.15.3 An overview of methods for setting cut-off scores

In the following section, Hibpshman (2004:9-10) and Morgan and Michaelides (2005) have summarized a selection of common methods of setting cut-off scores.

2.15.3.1 The Angoff method

The Angoff method works by assembling a panel of experts for each test, presenting test items for their consideration, and asking them to estimate the proportion of persons with minimally acceptable skills in the given content area who would be expected to get each item right. A criterion is set in advance for the proportion of items that must be judged job relevant in order for the test to be

deemed a valid measure of performance. After all the items have been rated, the judgments of the raters are combined to determine a cut-off score for the whole test. The assumption behind the use of this method is that the testers desire to assure that test takers have some minimum level of content or basic skills knowledge. The standard error can also be calculated for the cut-off score. A lower standard error is desirable since it denotes better agreement among the judges.

2.15.3.2 Modified Angoff Method

In the modified Angoff method, judges are asked to picture a hypothetical borderline examinee and indicate the probability (between 0.00-1.00) that a candidate will correctly answer each test item. Another way to consider this task is to picture 100 borderline candidates and determine how many of them would answer the item correctly. These probabilities are summed for each judge to determine each individual judge's cut-off score. Then, the individual cut-off scores are averaged across all judges to obtain the recommended cut-off scores. This method works well for tests with dichotomously scored items, and has been used in assessments that are primarily multiple-choice but also include some open-ended items.

Like all methods, the Modified Angoff includes multiple rounds of ratings accompanied by judge discussion between rounds. This method has been well researched and has a long precedence. Another advantage is that it does not require candidate data (other than impact data) be present, which makes it less vulnerable to time constrictions. A criticism is that it may be difficult for judges to accurately assign probabilities across the range from 0.00 to 1.00. This may result in only a few probability values being used, and depending on discrepancies between judges, there may be a lack of internal consistency. Another potential drawback is that judges may lose sight of the candidates' overall performance on the assessment due to the focus on individual items.

2.15.3.3 Body of work

In the Body of Work method (Kahl, Crockett, DePascale, and Rindfleisch, 1994, 1995; Kingston, Kahl, Sweeney, and Bay, 2001, Cited in Morgan and

Michaelides, 2005), judges examine complete sets of a candidate's work, including responses to both dichotomously and polytomously scored items. Judges review each candidate booklet and sort it into a performance category based upon its match to the Performance-Level Descriptors (PLDs). A small sample of candidate booklets across the range of possible scores is used as a range-finding activity to narrow down the approximate locations for where the cut scores should be placed. Using the defined range, sample candidate booklets are chosen to represent every score point between the lowest possible score in the range and the highest possible score in the range. Judges are then asked to work on one cut-off score at a time and sort booklets into one of the two performance categories surrounding the cut-off score. The test scores where a candidate is equally likely to belong to either group as determined by logistic regression are used to identify the final cut-score placements. An advantage of the Body of Work method is the relatively simple task of assigning candidate booklets to performance groups and the fact that judges are working with real candidate responses. A criticism is the amount of preparation time and the need for large quantities of candidate work available from which to pull the pinpointing round examples at every score point under consideration. However, this is a solid method for tests that are primarily performance based.

2.15.3.4 Bookmark

In the Bookmark method (Lewis, Mitzel, and Green, 1996; Mitzel, Lewis, Patz, and Green, 2001, Cited in Morgan and Michaelides, 2005), test items are ordered from easiest to most difficult based on Item-Response Theory (IRT) b-values, difficulty parameters, or some other index of item location. Judges are asked to consider items in the order of difficulty and identify the place in the ordered item booklet where the borderline candidate at each performance category would have a specific response probability (RP), traditionally two-thirds (RP67), of getting the item correct. Judges are instructed to place a bookmark into the ordered item booklet at the identified spot to mark their recommended placement for the cut-off score. After three rounds of bookmark placement with discussion between each round, final-round judges' bookmark placements are compiled and the median is selected for the cut-off score recommendation. This cut-off score recommendation is then located on the IRT ability metric to find the place where students have a two-thirds

(or other probability being used) chance of answering the identified item correctly and this becomes the final cut-score recommendation. Recent modifications to the Bookmark method include using small discussion groups between the rounds to diminish the influence of one strong judge and asking judges to work as a group to determine what each item measures and what makes it difficult prior to setting the first bookmark. An advantage of the Bookmark method is the ability to set multiple cut-off scores simultaneously. The method is also very efficient in terms of time needed and seems to be easily understood by judges. This method works well with both dichotomously and polytomously scored items. A criticism is the use of the RP67 value, which is arbitrary and can be confusing to judges and authoritative bodies who think the judges' bookmark placements are directly translated as the recommended cut score. The Bookmark method is one of the most widely used cut-off score methods in recent years.

2.15.3.5 Borderline Group

The Borderline Group method (Livingston and Zieky, 1982; Zieky and Livingston, 1977, Cited in Morgan and Michaelides, 2005) relies on the identification of a group of examinees as "borderline." Judges categorize examinees with whom they are familiar as adequate, inadequate, or borderline. This categorization is based on their evaluations of the examinees' proficiencies and their understanding of borderline performance on the skills being assessed, but without any consideration of the examinees' actual performance on the test. When the borderline examinees are selected, the median of their scores on the assessment is defined as the cut-off score. It is a very simple method to use and explain, although it may be difficult to identify students who are truly "borderline". The judges make decisions about their own students regarding the students' proficiency in the domain being assessed. Group membership decisions should be made based on performance information and free of irrelevant information that may consciously or unconsciously influence the judges' opinions, such as attendance or personality. This unbiased categorization may be difficult to accomplish and is one of the criticisms of this method.

2.15.3.6 Contrasting Groups methods

In the Contrasting Groups method (Bingham, 1937; Livingston and Zeiky, 1982; Nedelsky, 1954, cited in Morgan and Michaelides, 2005), instructors who are familiar with the students taking the test study the Performance-Level Descriptors (PLDs) and then categorize each of their students into one of the performance levels. Tests administered to the groups are scored, and score distributions are produced. The score distributions for each group (e.g., those students classified as Entry-Level Course and those classified as Advanced-Level Course) are plotted and the cut-off score is identified as the point at which the two distribution curves intersect. An alternative is to select as a passing score, the score that results in the fewest false positive and false negative classifications (Sireci, Robin, and Patelis, 1999). Webb and Miller (1995) used a variation of the Contrasting Groups method where raters reviewed papers written in response to constructed response items and sorted the existing papers, rather than students, into categories. An advantage of this method is the ability to accommodate both dichotomously scored and polytomously scored items. An additional advantage is the ability to collect data prior to the administration of the exam. Contrasting Groups is considered a good method to use when revisiting cut-off score decisions to provide confirmatory evidence that the decisions are still valid (or evidence of the need to run a new cut-score study). A disadvantage of this method is that it can be subject to how well raters know students being classified and any personal feelings they have toward those students.

To sum up, a variety of standard-setting methods have been developed. However, many of the methods work best with a particular item type, and thus matching the test format to an appropriate method should help determine which standard-setting method should be used. Because it is possible that different standard-setting methods may result in different recommended cut-off scores, it is essential that careful thought goes into the decision of which standard-setting method is to be used.

To help the test developers in making decisions on which standard-setting method to use, general steps in typical process for setting cut-off scores are presented in the following paragraph.

2.15.4 Common Steps in a Standard –Setting

While each standard-setting method has its own set of unique steps or features, Morgan and Michaelides (2005:2) interestingly suggest a general 12 steps in the typical process for setting cut-off scores. The 12 general steps are listed below:

1. Identify the purpose and goals of the cut-off score study.
2. Choose an appropriate method for setting cut-off scores.
3. Choose a panel of subject-matter experts and stakeholders to participate.
4. Write Performance-Level Descriptors (PLDs).
5. Train the panelists on the selected cut-off score method.
6. Train the panelists on the content standards and assessment(s) to which the cut-off score will be applied.
7. Compile item ratings or holistic judgments from the panelists that can be used to calculate cut-off score(s).
8. Conduct panel discussions regarding the judgments and resulting cut-off score(s).
9. Present consequences or impact data to the panel (optional).
10. Conduct a panelist evaluation of the process and their level of confidence in the resulting cut-off score(s).
11. Compile technical documentation to support the validity of the process for setting cut-off score(s). Make recommendations to college administrators.
12. College administrators make the final decision.

Documenting the process for validity purposes starts with the very first step. Not only is it important to keep a record of the content standards, PLDs, rosters of committee members, and data recording sheets, it is necessary to document all decisions as well. These decisions include determining the number of cut-off scores, selecting a method, choosing the panel, writing the PLDs, training the panelists, determining the feedback given, and calculating the cut-off scores. These steps

should be documented first in a plan of action, and then again in a final technical report. It should be noted that an integral part of the validity for any cut-off score process is ensuring that the testing instrument is appropriate for the student population and the intended purpose of course placement.

2.15.5 The Use of cut-off scores

This section summarizes papers and studies on the use of cut-off scores in many aspects.

2.15.5.1 Setting Cut-off scores in professional licensure tests:

Hibpshman (2004) mentions in his paper entitled " Considerations related to setting cut-off scores for teacher tests" that all states in America require licensure tests for teacher certification as they were seen as a means of raising the quality of teaching by assuring that teachers had minimum levels of literacy and context knowledge. The use of tests for these purposes began and spread quickly nationwide.

He further explains that the states having a right to require certification of teachers is rarely disputed, but particular requirements related to certification, including testing, are sometimes controversial. Testing by its nature must distinguish between groups of individuals, those who pass and those who do not. And persons who fail often believe that the test did not fairly measure their ability. Thus, the process of determining cut scores should not be done capriciously and that the testers have reasoned and professionally defensible rationale for the levels they select.

Hibpshman (2004:7) emphasizes in his paper that test construction as established in the standards requires that two features of a test be established, reliability and validity. Reliability means that a test score should be a consistent measure of some trait. Validity means that a test should have proven value for some particular purpose (Hibpshman , 2004).

Two matters relating to reliability are of importance. First, reliability is an essential precondition for establishing validity: mathematically, a test's validity is

bound by its reliability. Secondly, in an index derived from reliability studies, the standard error of measurement (SEM) is of important in setting and evaluating cut-off scores. The SEM is a measure of the amount of uncertainty in a test score that describes a region into which an individual's true score can be expected to fall. This becomes important in setting cut- off scores since wherever we set the cut-off score, there will be a number of individuals who might otherwise pass the test, but fail to do so for reasons not related to the trait measured by the test. Of course there are also persons who should really have failed the test, but pass for reasons having nothing to do with the trait of interest. These two types of misclassification are known as false negatives and false positives. In principle, both types of misclassification must be minimized, but in practice there is often a tradeoff between them. This tradeoff is often a consideration in the process of setting cut-off scores.

For test validity, Hibpshman (2004) mentions that three different types of evidence have been used to make inferences about test validity. They are content validity, criterion-related validity and construct validity. Among these three, criterion-related validity is virtually impossible to establish in many cases because adequate measures of performance are difficult to obtain, and sometimes it is difficult to arrive at an operational definition of good performance. It is widely believed that academic proficiency and content knowledge are essential in determining who will be a good teacher, but in fact research studies provide weak support for this idea. However, it was reported that the Angoff method, developed in 1971 by William H. Angoff, is the method used in setting the cut-off scores for teacher tests.

2.15.5.2 Setting cut-off scores for Knowledge Tests Used in Promotion, Training, Certification, and Licensing.

Biddle (1993) has proposed a very interesting method in setting job related cutoff scores in his paper, "How to set cut-off scores for knowledge tests used in promotion, training, certification, and licensing."

The suggested cut-off score setting process incorporates the advantages of job related process reviewed by the United Supreme Court, adds some job related

features to it, then combines the modified job related process with a distribution-wide adverse impact analysis. This process is called "Unmodified Angoff". In this process, seven to ten subject matter experts are used to give input on the job relatedness of a test and its items, at least 50 percent of the Subject Matter Experts need to agree on issues that determine inclusion of an item in a test. A higher standard would be 70 percent Subject Matter Expert agreement. Then ask the Subject Matter Expert to answer a question on the job relatedness of the item. A preferred option would be to have the Subject Matter Experts identify the duty(s) for which the knowledge measured by the test item is needed to competently perform the duty(s). The Subject Matter Experts are also asked to identify the level of consequence for what could likely happen in terms of duty performance if a person performing a duty and needing the knowledge measured by the test item does not know the answer to the item. After that each Subject Matter Expert states the probability for the minimally acceptable score of items that a person would answer correctly. These probabilities are summed up and the final average representing the average minimally acceptable score is identified. Finally, the reliability is calculated before the standard decision of the test is made. Also, statistical and human factors such as the size of standard error of measurement, risk of error, internal consistency of the Angoff subject matter expert panel, supply and demand for the jobs in the work force should be considered. After this consideration, the Angoff average score by one, two, or three standard errors of measurement is adjusted.

2.15.5.3 Setting cut-off scores on large scale assessment

Claycomb (1999) reported in her paper entitled " Setting Cut-off Scores on Large Scale Assessment" that although standardized tests have been used by schools and districts to evaluate and sort students for almost a century, what is new about today's assessment programs is their sophistication, variety, and emphasis on standards.

If standards are the skeleton around which states build their education systems and assessments are the muscles that bring the system to life, then cut-off scores are the vital signs by which the quality of life is evaluated. Cut-off scores, the actual numerical limits, are applied to student performance on an assessment. These

scores define how well students, teachers, schools, or the education system are performing. They tie student performance on an assessment directly to the standards, and fundamentally define at least two important components of a standards-based system. In other words, they provide a clear measure of "how good is good enough" and provide a standard yardstick by which to measure progress.

Claycomb (1999: 3) further explains that there are several key ways in which good cut-off scores support the success of state standards and assessment systems:

- Cut-off scores provide a system to measure existing levels of performance.
- Cut-off scores provide a yardstick by which policymakers can set future achievement goals.
- Cut-off scores, and the scoring rubrics associated with them, provide students and teachers with actual examples of the kind of work expected of them under the new standards.
- Cut-off scores provide a yardstick by which assessment results can be understood.

2.15.5.4 Setting cut-off scores on placement

Morgan and Michaelides (2005) state in their research report that course placement decisions for students entering college can have a significant impact on student's academic preparation and time they will spend in college before completing a degree. Students who can begin their studies at more advanced level courses, resulting from successful placement test results are able to take advantage of more advanced courses and complete their degree requirements early. In contrast, the students who are unable to show proficiency in college placement tests may be required to complete remedial courses until sufficient proficiency is gained.

Due to this high stakes decision that may be attached to the placement test, it is important that the placement process be as solid and defendable as possible. The use of cut-off sores that classify the students into categories is an integral part of the placement process. Their research report helps college administrators make valid decisions regarding setting cut-off scores.

In Thailand, Teo and Chatupote (2003) conducted research on how to establish the cut-off scores for placing first year students in required English courses at Prince of Songkla University (PSU) Hat Yai Campus. In the results of their study, four cut-off scores were established based on scores of the English Entrance Examination.

2.15.6 Selecting Method for Setting Cut-off Scores

The number of cut-off score methods increases each year. Currently, there are at least 50 methods that test developers and measurement experts use to set assessment cut-off score (Berk, 1995 cited in Claycom, 1999). This variety arises due to the fact that, although testing experts may prefer certain methods and researchers may point out ways in which methods differ, no single method is universally best or, more accurate. Furthermore, there is no "gold standard" (Claycomb, 1999: 4) to which the results of different cut-off score methods can be compared, so it is impossible to tell which cut score is actually closest to the real standard.

As stated in Claycomb (1999) and Morgan and Michaelides (2005), the common methods currently in use to set cut scores for an assessment system can be described in one of three ways: test-centered, examinee-centered, or combined (compromise) methods.

Test-centered methods are those in which the judges focus solely on the test content and/or item level information, while examinee-centered methods require that judges examine the students' performance more holistically. Compromise methods employ both absolute and normative standards to set cut-off scores.

Until recently, most assessment cut-off scores were set using test-centered methods in the large part because test-centered methods are particularly amenable to multiple-choice questions. Only recently, particularly as a result of the growth of new testing methods that include performance evaluations, have examinee-centered methods come to have been used frequently as a test-centered one.

Regardless of which method is utilized to set cut-off scores, it is important to recognize that every method has strengths and weaknesses. It is therefore difficult to draw concrete conclusions about which method is better or worse. Generally, the method of setting standards depends upon the type of test and its intended use. Moreover, it is important to concentrate not only on the way in which cut-off scores are set, but once they have been set, one should consider those cut-off scores in terms of their defensibility, validity, reliability, fairness, and political acceptability (Claycomb, 1999:5).

2.15.7 Evaluating the Process and Standards

Reliability and validity of standard setting

Regardless of the specific method used to set cut scores, every method needs to be checked for validity and reliability. In general, validity measures ensure that cut scores really represent the intended standards. Reliability measures, on the other hand, establish that judges' decisions about where to set cut-off scores are consistent and replicable. Both measures are indispensable in building a system of standard-based assessment that is fair and credible. As suggested in Claycomb (1999:7-8), in general, there are at least four different measures that testers should look for in order to ensure that the cut scores they apply to individuals taking the test are valid and reliable.

Checking cut-off score setting methods for validity and reliability is not the same thing as establishing validity and reliability for assessment. Throughout the cut-off score process, consideration should be given to the type of documentation that should be maintained. Documentation includes the plan for the cut-off score study; any scripts used; the materials given to panelists; any slide show presentation given; panelists' ratings; panelists' evaluations of the process and the resultant cut-off scores; the impact data that was presented to the panelists; and data used to create any other materials used in the cut-off score session, such as score distributions and any item-difficulty estimates that may have been used for item ordering. The documentation provides evidence to support the validity of the cut-off scores. Kane

(1994) provides two guidelines for examining the validity of performance standards: 1) that the cut-off score corresponds to the specified performance standard and 2) that the specified performance standard is reasonable given the purpose of the decision.

It is also standard procedure to create a technical report following the cut-off score session that describes the procedures and summarizes panelists' ratings and evaluations, as well as a summary of panelists' comments provided on the evaluation forms. The technical report should summarize the impact data; provide the standard errors of judgment (SEJs) for each cut-off score and the standard error of measurement (SEMs) for the test. It is good practice to provide the final cut-off score recommendations along with values representing +/-2 SEJs and +/-2 SEMs. Along with the resultant cuts, it is helpful to provide estimates of the percentages of students in each performance category based on the cut-off scores +/-2 SEJs and +/-2 SEMs for the total population and possibly for any subgroups of interest (Morgan and Michaelides, 2005:7)

2.15.8 Writing Performance-Level Descriptors

An initial step to set cut-off scores is the creation of Performance Level Descriptors (PLDs) or working definitions of each of the performance levels. The PLDs describe the meaning behind words like "basic," "proficient" and "advanced" or clearly delineate the difference in expectations for students in a remedial course, an entry-level course, or an advanced course in the subject area.

Many raters have reported that when they are thinking of what it means to classify a student into categories of proficient and advanced or into categories delineated by course level, they will often picture a student from their class whom they feel would be classified into that performance category. This can be extremely useful in helping the panelists fully conceptualize the task of setting cut-off scores. However, it would not be realistic to expect that all panelists come into the cut-off score session with the same student in mind for meeting the requirements to be placed into a particular course. Therefore, it is necessary to calibrate the raters through discussions of the content standards and the degree to which the standards

must be mastered for a student to be classified into each performance level. The creation or refinement of PLDs facilitates the calibration of panelists by providing each rater with the same working definition for each performance level. The PLDs may be created during the process of setting cut-off scores. However, to reduce requirements for rater time, an alternative is to convene a panel of experts for the purpose of creating the PLDs prior to the cut-off score session. Then, during the process of setting cut-off scores, raters are given the prepared PLDs and provided an opportunity to discuss, edit, and refine them. The process of setting cut-off scores should not proceed until the PLDs are to the point that the raters feel comfortable that they reflect what students at each performance level should know and be able to do. It is essential for all cut-off score methods that the individual members of the raters have the same understanding of the performance levels, and that they are specifically focusing on the definitions at the borderline level or the "just sufficiently knowledgeable" student. That is, they know what it means to be just barely proficient enough for the entry-level course or just barely advanced enough for a non-entry-level or subsequent course.

Regardless of the process used to produce the final working definition, the PLDs should:

- Describe what test takers at each level should reasonably know and be able to do.
- Relate directly to the content standards, course prerequisites, and course requirements.
- Distinguish clearly from one level to the next.
- Be written in positive terms.
- Be written in clear and concise language without using non-measurable qualifiers such as often, seldom, thorough, frequently, limited, etc.
- Focus on achievement.

2.16 Proficiency scales and guidelines of related standard tests

Standardized Proficiency Scales in language assessment have been widely established and used among educators, students, testers, test takers, and related stakeholders. As the aim of this study is to investigate Thai students' listening ability

in English for the service and hospitality industry, the researcher will take only two standardized proficiency scales and guidelines into consideration in order to write descriptions for the established proficiency level. These include scales and guidelines of American Council on Teaching of Foreign Languages (ACTFL) and the Test of English for International Communication (TOEIC).

The proficiency scales and guidelines of the ACTFL are referred to in writing ability descriptors for the L-PESH Test because the ACTFL has become synonymous with innovation, quality, and reliability in meeting the changing needs of foreign language educators and their students. Moreover, the ACTFL is widely accepted and referred to by teachers of all languages at all educational levels (American Council for the Teaching of Foreign Languages, 1983).

For the Test of English for International Communication (TOEIC), the researcher refers to it because most of the time, graduating students in the field of service and hospitality are required to take the TOEIC Test and present their TOEIC scores to the employers when they apply for a position or want to get promoted. Moreover, the format of the L-PESH Test is similar to the TOEIC Test. While the TOEIC Test measures communicative English in general, the L-PESH Test measures more specific English in service and hospitality industry.

2.16.1 The American Council on the Teaching of Foreign Languages (ACTFL)

The American Council the Teaching of Foreign Languages (ACTFL) is the only national organization dedicated to the improvement and expansion of the teaching and learning of all languages at all levels of instructions. ACTFL is an individual membership organization of more than 7,000 foreign language educators and administrators from elementary through graduate education, as well as government and industry (American Council for the Teaching of Foreign Languages, 1983).

Over the past 30 years, the ACTFL has become synonymous with innovation, quality, and reliability in meeting the changing needs of foreign language educators

and their students. From the development of Proficiency Guidelines, to its leadership role in the creation of national standards, ACTFL focuses on issues that are critical to the growth of both the professional and the individual teacher. Through their membership, new as well as veteran teachers are making an important investment in the future. The ACTFL was founded in 1967 by the Modern Language Association of America. It remains the only national organization representing teachers of all languages at all educational levels (American Council for the Teaching of Foreign Languages, 1983).

ACTFL proficiency guidelines

This section contains descriptions of different levels of language proficiency identified by the American Council the Teaching of Foreign Languages based on the five levels originally defined by the US Foreign Service Institution. These descriptions can be helpful in setting language-learning goals, in planning learning activities and evaluation proficiency. In this study, the aim of the L-PESH Test is to measure listening ability of the test takers; therefore, only detailed description on listening skills is presented in the following paragraphs.

ACTFL guidelines: Listening

According to ACTFL Proficiency Guidelines (revised in 1985), the description is divided into five levels as follows:

ACTFL guidelines: Novice

Novice-Low

Understanding is limited to occasional isolated words, such as cognates, borrowed words, and high-frequency social conventions. Essentially there is no ability to comprehend even short utterances.

Novice-Mid

Able to understand some short, learned utterances, particularly where context strongly supports understanding and speech is clearly audible. Comprehends some words and phrases from simple questions, statements, high-frequency commands and courtesy formulae about topics that refer to basic personal information or the immediate physical setting. The listener requires long pauses for assimilation and periodically requests repetition and/or a slower rate of speech.

Novice-High

Able to understand short, learned utterances and some sentence-length utterances, particularly where context strongly supports understanding and speech is clearly audible. Comprehends words and phrases from simple questions, statements, high-frequency commands, and courtesy formulae. May require repetition, rephrasing, and/or a slowed rate of speech for comprehension.

ACTFL guidelines: Listening--Intermediate

Intermediate-Low

Able to understand sentence-length utterances that consist of recombination of learned elements in a limited number of content areas, particularly if strongly supported by the situational context. Content refers to basic personal background and needs, social conventions and routine tasks, such as getting meals and receiving simple instructions and directions. Listening tasks pertain primarily to spontaneous face-to-face conversations. Understanding is often uneven; repetitious and rewording may be necessary. Misunderstandings in both main ideas and details arise frequently.

Intermediate-Mid

Able to understand sentence-length utterances that consist of recombination of learned utterances on a variety of topics. Content continues to refer primarily to basic personal background and needs, social conventions and somewhat more

complex tasks, such as lodging, transportation, and shopping. Additional content areas include some personal interests and activities, and a greater diversity of instructions and directions. Listening tasks not only pertain to spontaneous face-to-face conversations but also to short routine telephone conversations and some deliberate speech, such as simple announcements and reports over the media. Understanding continues to be uneven.

Intermediate-High

Able to sustain understanding over longer stretches of connected discourse on a number of topics pertaining to different times and places; however, understanding is inconsistent due to failure to grasp main ideas and/or details. Thus, while topics do not differ significantly from those of an Advanced level listener, comprehension is less in quantity and poorer in quality.

ACTFL guidelines: Listening--Advanced

Advanced

Able to understand main ideas and most details of connected discourse on a variety of topics beyond the immediacy of the situation. Comprehension may be uneven due to a variety of linguistic and extra linguistic factors, among which topic familiarity is very prominent. These texts frequently involve description and narration in different time frames or aspects, such as present, non-past, habitual, or imperfective. Texts may include interviews, short lectures on familiar topics, and news items and reports primarily dealing with factual information. Listener is aware of cohesive devices but may not be able to use them to follow the sequence of thought in an oral text.

Advanced Plus

Able to understand the main ideas of most speech in a standard dialect; however, the listener may not be able to sustain comprehension in extended discourse which is propositionally and linguistically complex. Listener shows an

emerging awareness of culturally implied meanings beyond the surface meanings of the text but may fail to grasp sociocultural nuances of the message.

ACTFL guidelines: Listening--Superior

Superior

Able to understand the main ideas of all speech in a standard dialect, including technical discussion in a field of specialization. Can follow the essentials of extended discourse which is propositionally and linguistically complex, as in academic/professional settings, in lectures, speeches, and reports. Listener shows some appreciation of aesthetic norms of target language, of idioms, colloquialisms, and register shifting. Able to make inferences within the cultural framework of the target language. Understanding is aided by an awareness of the underlying organizational structure of the oral text and includes sensitivity for its social and cultural references and its affective overtones. Rarely misunderstands but may not understand excessively rapid, highly colloquial speech or speech that has strong cultural references.

ACTFL guidelines: Listening--Distinguished

Distinguished

Able to understand all forms and styles of speech pertinent to personal, social, and professional needs tailored to different audiences. Shows strong sensitivity to social and cultural references and aesthetic norms by processing language from within the cultural framework. Texts include theater plays, screen productions, editorials, symposia, academic debates, public policy statements, literary readings, and most jokes and puns. May have difficulty with some dialects and slang.

2.16.2 The TOEIC (Test of English for International Communication)

The following sections present information about the TOEIC. This information was drawn from "TOEIC From A to Z" (2003).

The Test of English for International Communication - TOEIC, is an English language proficiency test for people whose native language is not English. It measures the everyday English skills of people working in an international environment. TOEIC test scores indicate how well people can communicate in English with others in the global workplace. The test does not require specialized knowledge or vocabulary beyond that of a person who uses English in everyday work activities. Today, the TOEIC test has become the world's leading test of English language proficiency in a workplace context. More than 8,000 corporations worldwide have used the TOEIC test and more than 2 million people take the test every year.

The TOEIC test is used primarily in the workplace. A wide range of companies, from small businesses to multinationals to government agencies, operating in many different industries and countries use the test. The TOEIC test is an important management tool that allows organizations to make significant personnel decisions.

The TOEIC test consists of 200 multiple-choice questions; 100 listening comprehension questions, and 100 reading comprehension questions. The listening comprehension section is administered by audiotape; the reading comprehension section is administered using a standard paper-and-pencil format. The answers from both sections are recorded on a scan-able answer sheet. Examinees receive two sub scores, one each for listening comprehension and reading comprehension, along with a total score (listening comprehension plus reading comprehension). Each standardized sub score ranges from 5 to 495, with a total score range of 10 to 990.

Test scores can give TOEIC users very general information about a test-taker. However, a score itself does not provide information about an examinee's specific English-language abilities: It does not provide information about the specific actions or behaviour that an examinee can perform or may be expected to perform in English. For example, a score does not provide information about what an examinee with a total score of 400 may be able to do in English as compared to an examinee with a total score of 300. Furthermore, the scores do not differentiate between candidates in different score ranges (for example 200 to 250, 300 to 350) in terms of English use.

Uses of the TOEIC in Thailand

The Test of English for International Communication (TOEIC) is a test of listening and reading proficiency, first developed by the Educational Testing Service, Princeton, in 1979. Today the TOEIC test is the world's most widely used English language proficiency test, with close to 3 million tests administered worldwide annually. The TOEIC test was first administered in Thailand in 1988. The first client, The Regent of Bangkok, remains a TOEIC client today. As a broad range test of English language, the TOEIC test is used by different organizations for different purposes.
(http://www.toeic.co.th/TOEIC/Htmls/Uses.html#anchor530513#anchor530513).

The TOEIC Test and Recruitment

Once an organization has identified and set the language standards it needs for English-essential positions, the TOEIC test can then be used as an integrated component of the corporate recruitment process.

Today's Human Resources departments need as much information about each potential candidate as possible, to make the best hiring decision they can. In the current economic environment organizations need to be sure that they are hiring staff with the best combination of skills and knowledge, to minimize the need for company sponsored training to bring them up to required standards.

The TOEIC test has successfully been incorporated into many corporations' recruitment procedures in a variety of ways.

Benchmarking in TOEIC

Benchmarking is a procedure that compares the TOEIC test scores of groups of people within an organization whose ability to perform their jobs using English is known. Clients are subsequently able to develop a series of English proficiency levels that can be applied to different jobs within the company and against which employees can be evaluated. The TOEIC service encourages all institutional users of the test to follow this procedure whenever possible. Following is the table presenting five levels of English competency and their descriptors.

Table 2.2 Five Levels of English Competency and Descriptors

TOEIC Score	Description
900-990	Managers who are able to represent the company unaccompanied and with final authority in negotiating agreements and contracts with native English-speaking partner organizations.
800-850	Managers who are able to represent the company unaccompanied in contributing to the negotiation of agreements and contracts with partner organizations using English.
700-750	Individuals who actively participate in meetings with partner organizations using English
600	Individuals who accompany and support staff members with primary responsibility for business meetings. May be called upon to give a short, prepared speech and/or to take the minutes of the meeting.
400-450	Individuals who, with the assistance of vocabulary/grammar aids have occasional and short-term contact in English. This may include welcoming visitors (in person or by telephone) and working with the mail.

(http://www.ets.org/Media/Tests/Test_ofEnglish_for_International_Communication/TOEICAZ.pdf)

These five descriptors may be used as guidelines only. They should be adapted to real situations and should not be considered definite.

TOEIC Can-Do Guide

From the review of the "Can-Do Guide Linking TOEIC Scores to Activities Performed Using English" (2000), the researcher finds information that allows users of the TOEIC test to link TOEIC scores to the activities that examinees may or may not be able to do in English. The tables in the guide provide examples of the activities that examinees are likely to be able to perform in English given certain Reading Comprehension scores and Listening Comprehension scores.

As this research focuses only on listening ability, the following TOEIC Can-Do tables focus on presenting the linkage of listening scores to activities performed using English.

Table 2.3 Can-Do Guide for TOEIC Listening Score of 5 – 100

Can do	
Can do with difficulty	♦ understand simple questions in social situations such as "How are you?" "Where do you live?" and "How do you feel?" ♦ understand a salesperson when she or he tells me prices of various items ♦ understand someone speaking slowly and deliberately, who is giving me directions on how to walk to a nearby location
Cannot do	♦ understand explanations about how to perform a routine task related to my job ♦ understand a co-worker discussing a simple problem that arose at work ♦ understand announcements at a railway station indicating the track my train is on and the time it is scheduled to leave ♦ understand headline news broadcasts on the radio ♦ understand a client's request made on the telephone for one of my company's major products or services ♦ understand a person's name when she or he gives it to me over the telephone ♦ understand play-by-play descriptions on the radio of sports events that I like (e.g., soccer, baseball) ♦ understand an explanation given over the radio of why a road has been temporarily closed ♦ understand someone who is speaking slowly and deliberately about his or her hobbies, interests, and plans for the weekend ♦ understand directions about what time to come to a meeting and the room in which it will be held ♦ understand a discussion of current events taking place among a group of persons speaking English ♦ understand an explanation of why one restaurant is

Table 2.4 Can-Do Guide for TOEIC Listening Score of 105-225

Can do	
Can do with difficulty	♦ understand simple questions in social situations such as "How are you?" "Where do you live?" and "How do you feel?" ♦ understand a salesperson when she or he tells me prices of various items ♦ understand someone speaking slowly and deliberately, who is giving me directions on how to walk to a nearby location ♦ understand a person's name when she or he gives it to me over the telephone understand directions about what time to come to a meeting and the room in which it will be held
Cannot do	♦ understand explanations about how to perform a routine task related to my job ♦ understand a co-worker discussing a simple problem that arose at work ♦ understand announcements at a railway station indicating the track my train is on and the time it is scheduled to leave ♦ understand headline news broadcasts on the radio ♦ understand a client's request made on the telephone for one of my company's major products or services ♦ understand play-by-play descriptions on the radio of sports events that I like (e.g., soccer, baseball) ♦ understand an explanation given over the radio of why a road has been temporarily closed ♦ understand someone who is speaking slowly and deliberately about his or her hobbies, interests, and plans for the weekend ♦ understand a discussion of current events taking place among a group of persons speaking English ♦ understand an explanation of why one restaurant is better than another

Table 2.5 Can-Do Guide for TOEIC Listening Score of 230-350

Can do	
Can do with difficulty	♦ understand simple questions in social situations such as "How are you?" "Where do you live?" and "How do you feel?" ♦ understand a salesperson when she or he tells me prices of various items ♦ understand someone speaking slowly and deliberately, who is giving me directions on how to walk to a nearby location ♦ understand explanations about how to perform a routine task related to my job ♦ understand a co-worker discussing a simple problem that arose at work ♦ understand announcements at a railway station indicating the track my train is on and the time it is scheduled to leave ♦ understand headline news broadcasts on the radio ♦ understand a person's name when she or he gives it to me over the telephone ♦ understand someone who is speaking slowly and deliberately about his or her hobbies, interests, and plans for the weekend ♦ understand directions about what time to come to a meeting and the room in which it will be held ♦ understand an explanation of why one restaurant is better than another
Cannot do	♦ understand a client's request made on the telephone for one of my company's major products or services ♦ understand play-by-play descriptions on the radio of sports events that I like (e.g., soccer, baseball) ♦ understand an explanation given over the radio of why a road has been temporarily closed ♦ understand a discussion of current events taking place among a group of persons speaking English

Table 2.6 Can-Do Guide for TOEIC Listening Score of 355-425

Can do	♦ understand simple questions in social situations such as "How are you?" "Where do you live?" and "How do you feel?" ♦ understand a salesperson when she or he tells me prices of various items understand someone speaking slowly and deliberately, who is giving me directions on how to walk to a nearby location
Can do with difficulty	♦ understand explanations about how to perform a routine task related to my job ♦ understand a co-worker discussing a simple problem that arose at work ♦ understand announcements at a railway station indicating the track my train is on and the time it is scheduled to leave ♦ understand a client's request made on the telephone for one of my company's major products or services ♦ understand a person's name when she or he gives it to me over the telephone ♦ understand play-by-play descriptions on the radio of sports events that I like (e g . soccer, baseball) ♦ understand an explanation given over the radio of why a road has been temporarily closed ♦ understand someone who is speaking slowly and deliberately about his or her hobbies, interests, and plans for the weekend ♦ understand directions about what time to come to a meeting and the room in which it will be held ♦ understand an explanation of why one restaurant is better than another ♦ understand a discussion of current events taking place among a group of persons speaking English ♦ understand headline news broadcasts on the radio
Cannot do	

Table 2.7 Can-Do Guide for TOEIC Listening Score of 430-495

Can do	♦ understand simple questions in social situations such as "How are you?" "Where do you live?" and "How do you feel?" ♦ understand a salesperson when she or he tells me prices of various items ♦ understand someone speaking slowly and deliberately, who is giving me directions on how to walk to a nearby location ♦ understand explanations about how to perform a routine task related to my job ♦ understand announcements at a railway station indicating the track my train is on and the time it is scheduled to leave ♦ understand someone who is speaking slowly and deliberately about his or her hobbies, interests, and plans for the weekend ♦ understand directions about what time to come to a meeting and the room in which it will be held understand an explanation of why one restaurant is better than another
Can do with difficulty	♦ understand a co-worker discussing a simple problem that arose at work ♦ understand headline news broadcasts on the radio ♦ understand a client's request made on the telephone for one of my company's major products or services ♦ understand a person's name when she or he gives it to me over the telephone ♦ understand play-by-play descriptions on the radio of sports events that I like (e.g., soccer, baseball) ♦ understand an explanation given over the radio of why a road has been temporarily closed ♦ understand a discussion of current events taking place among a group of persons speaking English
Cannot do	

The following table, Table 2.8, shows the ability to use English in the workplace by using TOEIC scores and descriptions (on listening section). (http://www.toeic.ca/companies/TOEICresumescorefinalforweb.pdf)

Table 2.8 Ability to Use English in the Workplace by Using TOEIC Score

TOEIC score	Listening ability
455-495	Can: • understand native speakers of English in meetings • function in all of the situations described below whether professional or social, concerning concrete or abstract subjects
395-450	• understand most work related situations • understand most speakers of English in international meetings • function in all of the situations described below but with a greater degree of facility and accuracy
305-390	...understand: • explanations of work problems • requests for products on phone • discussions of current events by native speakers of English • headline news on radio
205-300	...understand: • explanations related to routine work tasks in one to one situations • some travel announcements • limited social conversations
130-200	• understand simple exchanges in everyday professional or personal life with a person used to speaking with non-native speakers • take simple phone messages
05-125	• understand adequately for immediate survival needs, directions, prices... • comprehend simple questions in social situations

In Thailand, many leading hotels have set their own requirements for TOEIC score considered essential for each position standard. The following table, Table2.8 presents the requirements of TOEIC score used in hotel staff recruitment. This

information was gathered from TOEIC requirements in the Pan Pacific Hotel, the Banyan Tree Hotel Bangkok, the Montien Riverside Hotel, and the J W Marriott Hotel.

Table 2.9 TOEIC-Essential Position Standards

Position	TOEIC Total	Score range
Accountant	625	700
Assistant Personnel & Administration Deputy Manager	700	700
Assistant to General manager	920	920
Cashier	525	525
Driver	0	175
Electrical Engineer; supervisor	550	600
Electrician	180	250
Engineer	650	700
Executive Director	650	800
Executive Secretary	800	920
Maid	175	525
Operator	450	525
Outlet Manager	850	850
Public Relation Officer	650	700
Purchasing Officer	650	750
Receptionist/ Operator	425	750
Sales and Marketing Manager	650	800
Translator	675	900
Waiter/Waitress	550	600

To sum up, as the aim of this study is to investigate Thai students' listening ability in English for the service and hospitality industry, the researcher will take only two standardized proficiency scales, the ACTFL and the TOEIC, into consideration in order to write descriptions of the proposed proficiency levels.

2.17 Conclusion and application in setting cut-off scores for the L-PESH Test

From the previous review of related literature, it can be concluded that to increase accountability among educators, students, testers, test takers, and related stakeholders, a test has to meet a certain standard which has been set. Students and test takers are expected to meet a standard of proficiency that the tests are designed to assess. This standard is later translated into a cut-off score in order to identify the level of ability of the test takers. The method used to set these cut-off scores is called standard setting.

The purpose of the cut-off score is to separate test takers who meet the standard from those who do not. It also identifies the levels of test takers' ability measured in the assessment. However, once the need to establish a performance standard, or to set cut-off scores, has been established, the question of what is the best method to be used in setting cut-off scores usually arises.

There are several acceptable ways to set standards. This variety arises due to the fact that, although testing experts may prefer certain methods and researchers may point out ways in which methods differ, no single method is universally best or, most accurate. No one standard-setting method is agreed upon as the best because it is possible that different standard-setting methods may result in different recommended cut-off scores.

Regardless of which method is utilized to set cut-off scores, it is important to recognize that every method has strengths and weaknesses. It is therefore difficult to draw concrete conclusions about which method is better or worse. Generally, the method of setting standards depends upon the type of test and its intended use. Moreover, it is important to concentrate not only on the way in which cut scores are set, but once they have been set, to consider those cut-off scores in terms of their defensibility, validity, reliability, fairness, and political acceptability (Morgan and Michaelides, 2005).

Cut-off scores can be used in many important assessments, such as setting cut-off scores in professional licensure tests, for knowledge tests used in promotion, training, certification, and licensing, on placement tests, and on large-scale assessment.

In Thailand there are few studies on setting cut-off scores and its use. However, from the literature review, it was found that setting cut-off scores was generally used in identifying students' ability in general English and in placement tests so as to place them into appropriate English courses. Moreover, studies on setting cut-off scores for English for Service and Hospitality assessment have not been found. Therefore, in this study the researcher decided to apply frameworks and methods suggested in the previous review to design her own practical method for setting cut-off scores for the new ESP test in this study, the L-PESH Test.

The type of cut-off scores of the L-PESH Test is a "Performance–related cut-off score". The method of setting these cut-off scores in this study is not based on any single method suggested above, but the researcher puts together some of suggestions from these methods and applies them to set cut-off scores for the L-PESH Test. The process of setting cut-off scores starts with having the students take the test, and applying descriptive statistics to analyze the test items and received scores. Next, ask the panel of experts to make decisions on the number of ability levels to be set. In this study, the experts and the researcher agreed to have eight ability levels for the test. This decision was based on the literature review on proficiency scales of related standard tests (ACTFL and TOEIC). Then, the cut-off scores are established by means of calculating the mean and the standard deviation of the L-PESH Test scores in the normal distribution. Finally, panel of experts is asked to consider the cut-off scores and to make final decision on appropriate cut-off scores.

For the interpretation of the cut-off scores, after the eight appropriate cut-off scores are set, they need to be interpreted. Therefore, the ability descriptors are elaborated by means of applying some suggestions from the literature review on the ACTFL proficiency guideline and TOEIC Can-Do guide. Since the L-PESH Test is an ESP test focusing on listening ability in English for the service and hospitality

industry, the proficiency descriptors describe what test takers at each level should know and be able to do. These descriptors relate directly to the content and objectives of the test and job market requirements. In addition, these descriptors can distinguish the test takers clearly from one level to the next.

2.18 Conclusion of this chapter

The review of literature in this chapter includes characteristics of hospitality language, the significance of listening ability in communication, listening proficiency, language proficiency tests, an overview of listening tests, factors affecting second language listening, approaches to assessing listening, implications for listening tests, English for Specific Purposes and its tests, a framework for analysing TLU and test task characteristics, essential components of LSP specifications, standard setting, identifying cut-off scores and their descriptors, and proficiency scales of related standardized tests. The researcher set out her research design and instruments based on this literature review.