CHAPTER III

REASEARCH METHODOLOGY

This chapter presents the research methods and the procedures of the study. Five major areas covered in this chapter are research procedures, subjects, research instrumentation, data collection, and data analysis.

3.1 Research procedures

Both quantitative and qualitative approaches were applied in this study. As the aim of the study is to investigate graduating students' listening ability in English for service and hospitality industry and to find the cut-off scores and ability descriptors for each level of listening ability, it is necessary that a new test developed to measure this ability is valid and reliable.

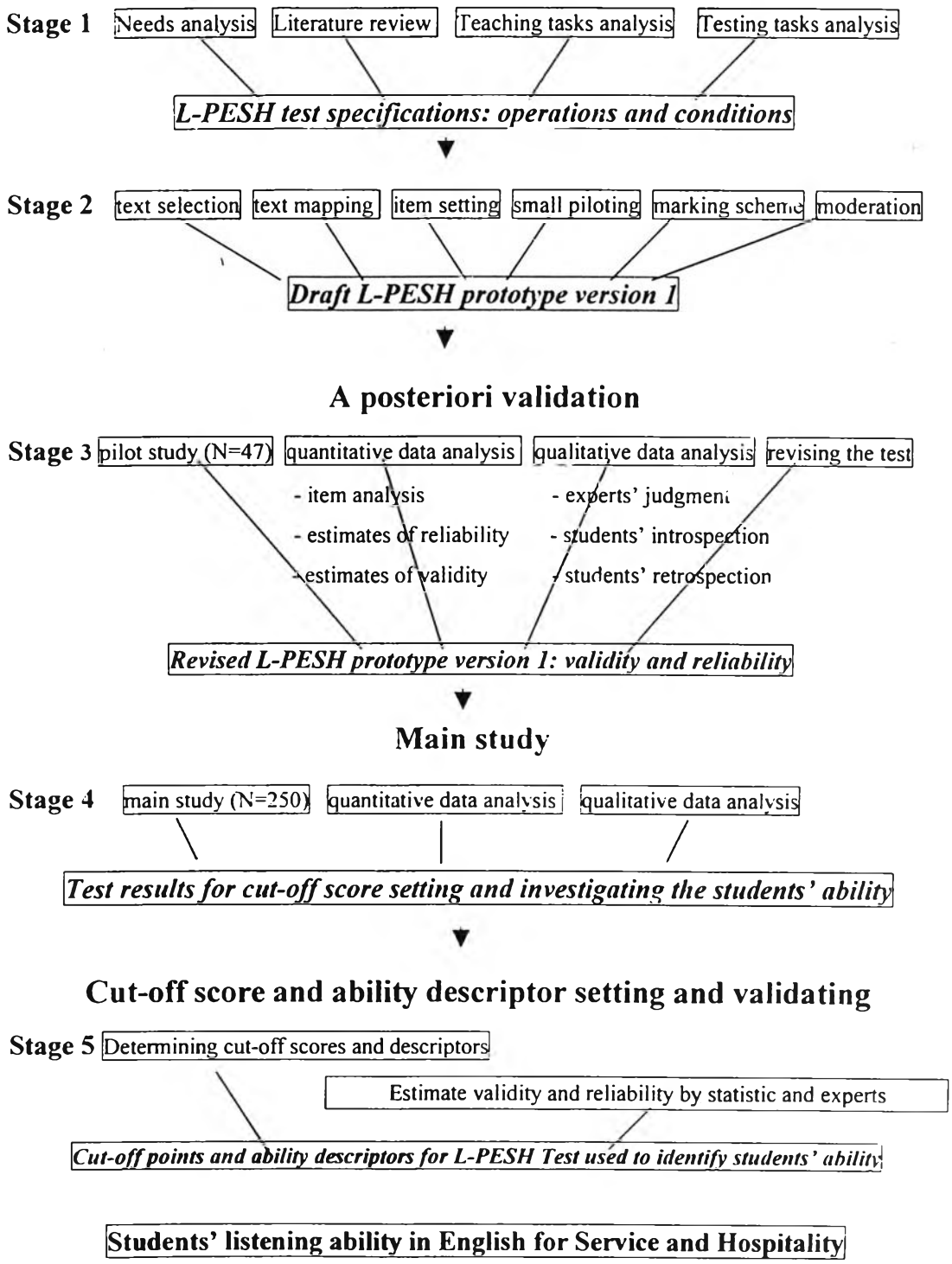
When a new test is being developed, there are usually two essential aspects to take into consideration: the test validity and the test reliability. The test writer has to ensure that the new test measures what it is supposed to measure, and the new test should provide consistency of measurement before the test result can be generalized to similar situations.

In constructing or selecting tests, the most important questions are: (1) to what extent can the interpretation of the scores be appropriate, meaningful, and useful for the intended application of the results? And (2) what are the consequences of the particular uses and interpretations that are made of the results?

The validity of the test is one of the basic concerns of language testing. It is generally considered desirable to establish validity in many ways as possible. Confidence in the test is directly proportional to the amount of evidence that is gathered in support of its validity.

Therefore, to develop a new test that is valid and reliable, the research procedure in this study is illustrated in the following diagram:

Figure 2: Stages in the research procedure



[Adapted from Urquhart and Weir (1998), presented in Weir et al (2000:5)]

3.2 Subjects

The subjects in this study were classified into two groups: the test taker group and the interviewee group.

3.2.1 The test taker group

This study applied the stratified random sampling technique to select the subjects. According to this technique, the population was subdivided into appropriate subgroups (strata) and then the subjects of each stratum were selected according to the predetermined sample size and purposes. In this study the researcher randomly selected four public and private universities in Bangkok that offer a course or a degree in the service and hospitality industry. The selected universities include Bangkok University, Kasem Bundit University, Kasetsart University and Rangsit University.

The test takers were 4th year Thai students studying in the departments of tourism and hotel industry at Bangkok University, Kasem Bundit University, Kasetsart University, and Rangsit University. These students have completed all required courses in English in their institutions. The sampling techniques were as follows:

3.2.1.1 Subjects for the pilot study: The 4th year students' name lists including their Grade Point Average (GPA) from two universities (Bangkok University and Kasetsart University) were obtained. Then the students were categorized into three groups; High, Medium, and Low, according to their Grade Point Average (GPA). Then 50 subjects from these three groups were randomly selected to be the test takers in the pilot study. The subjects in this group took two tests: the test of Listening Proficiency in English for Service and Hospitality (L-PESH Test and the Test of English for International Communication (TOEIC) on the same day.

3.2.1.2 Subjects for the main study: The students' name lists including their Grade Point Average (GPA) from four universities (Bangkok

University, Kasem Bundit University, Kasetsart University, and Rangsit University) were drawn the same way as in the pilot study. Next, the students from each university were divided into three sub-groups: High, Medium, and Low, according to their GPA. Then 60-65 subjects were randomly selected from each of these three sub-groups from each university. Finally, there were 250 subjects to be the test takers in the main study. These subjects took the revised version of the L-PESH Test.

3.2.2 The interviewee group: This group included HRD Managers, Training Managers, Department Heads, hotel staff working in selected hotels in Bangkok, and administrators and teachers working in those four selected universities.

A purposive sampling technique was applied with this group. The subjects consisted of 3 hotel personnel, 4 teachers and heads of departments, and 2 specialists in ESP test development, and 3 hotels guests. These subjects were interviewed both prior to and after the main study. The interviews prior to the main study focused on needs analysis related to the expected listening ability in English and other relevant qualification of the candidate. The interviews after the main study focused on justifying the cut-off scores and ability descriptors. This information was applied in developing the new test (L-PESH Test).

Table 3.1 illustrates the details of all subjects in this study.

Table 3.1 The Subjects for the Pilot and Main Studies

The studies	Total number	Details of the subjects
Subjects for the **pilot study,** as the test takers. They took two tests: the L-PESH test and the TOEIC test.	47	Graduating students, majoring in tourism and hotel industry. Twenty-four students from Bangkok University and Twenty- three students from Kasetsart. (These subjects were not included in the main study.)

Subjects for the **main study**, as the test takers. They took one test, the revised L-PESH test, only.	250	Graduating students, majoring in tourism and hotel industry. Sixty to sixty-five students from each of the following universities: Bangkok University, Kasem Bundit University, Kasetsart University, and Rangsit University.
Subjects for the **interviews**.	12	Three hotel personnel from selected hotels; the JW Marriott Hotel, the Pan Pacific Hotel, and the Monthien Reiversside Hotel, four Heads of Tourism and Hotel Industry Department and teachers from Bangkok University, Kasem Bundit University, Kasetsart University, and Rangsit University, two specialists in ESP test development, and three hotel guests; 1 Danish, 1 Japanese, and 1 American.

3.3 Research instrumentation

The following are the instruments used in this study.

3.3.1 Test of Listening Proficiency in English for Service and Hospitality Industry (L-PESH Test).

This new test was developed to be used as a tool for assessing the 4th year students' listening ability in English for the service and hospitality industry from four selected universities. Prior to the test specification development, a needs analysis had been conducted in order to gain essential information to be used in developing this new test. This need analysis included interviewing HRD managers, hotel personnel. Heads of the Department in selected universities, students, hotel guests and documentary analysis on textbooks and teaching materials. The information obtained from the needs analysis was applied in developing the test specification. In the first draft, 150 test items were written. Three specialists and one native speaker of English were invited to review the first draft of the test using the Item Objective Congruent test validating form (APPENDIX A). According to comments given by the specialists and the native speaker, the first draft of the test was edited and 26 test items were discarded. Later, the second draft of the test was ready for the pilot study. In this pilot version of the test, there were 124 test items, divided into four main parts including:

Part I:	Photographs	30 questions (4-choices)
Part II:	Question-response	35 questions (4-choices)
Part III:	Short conversations	33 questions (4-choices)
Part IV:	Short talks	26 questions (4-choices)
Total		124 questions
Total test time		1.30 hours

The test is a one-and-a-half hour, paper-and-pencil, and multiple-choice test.

The preparation of audiotape was conducted after the test review and revision. As the aim of the test is to measure listening ability in English for the service and hospitality industry, hence communication among participants is likely to involve a variety of Englishes. The speakers in the recording therefore included one

American native, one British native, one Japanese, one Chinese, two Swedish, and three Thai.

The test then was piloted with 47 test takers (randomly selected from Bangkok University and Kasetsart University). The test administration was held at Kasetsart University in the morning, while in the afternoon these 47 students also had to take a TOEIC test, administered by the TOEIC Thailand test centre.

In the pilot study, test results could be obtained from test administrations at Wittayaborigarn Building, room 310, Kasetsart University. The test takers were supposed to take both tests (the L-PESH Test and the TOEIC) in the test room on the same day. There were 47 selected test takers from two universities: Bangkok University and Kasetsart University. In order to avoid bias on test takers' ability in English, these test takers are purposively and randomly selected and classified into four groups according their GPA, Group One 2.00-2.50, Group Two 2.51-3.00, Group Three 3.01-3.50, and Group Four 3.51-4.00.

The test result was later analyzed in order to measure test validity and reliability.

3.3.1.1 Test validity: The validity is defined as the degree to which a test measures what it claims to be measuring. Validity was traditionally subdivided into three categories: content, criterion-related, and construct validity (Brown 1996:231-249).

a) Content validity To validate the content of the test, Osterlind (1998: 258) has explained that evidence for valid test-score interpretations is not inherent in the item-construction process but must be gathered through a systematic validation study. The procedures used for gathering content-related evidence for validity can be of great help to determine the quality of test items. The test writer can use the information found from this systematic study to examine and improve test items.

> *A content-validation study usually seeks to establish a consensus of informed opinions about the degree of congruence between particular test items and specific descriptions of the content domain that is intended to be assessed by those items. This typically requires convening a panel of expert judges who rate the item-to-content congruence according to some established criteria.*
>
> (Osterlind, 1998:258)

In addition, Hamp-Lyons, Hamilton, Lumley, and Lockwood (2003) have found from their study on "The Tester and the Specialist Informant" conducted in Hong Kong, that the language experts' perceptions of strengths and weaknesses of the texts were often not upheld by professional informants. What language experts see as an error may not be wrong in the specialists' perceptions.

Therefore, the content validity of the test used in this study was estimated by means of needs analysis, content analysis, and Target Language Use (TLU) analysis. In addition, three content specialists in the field were asked to review the test by applying the Item Objective Congruent Test Validating form. The form for test review was developed based on suggestions in Osterlind (1998) and Haladyna (1994). Finally, the Index of Item-Objective Congruence (IOC) was analyzed based on methods suggested in Brown (1996). See APPENDIX A for more details in IOC calculation.

The result of the IOC analysis showed that the average IOC Index of the L-PESH Test was 0.57. In addition, 73.39 % of the test items have the IOC index of ≥ 0.5. This means that 73.39 % of the test items with the IOC index of ≥ 0.5 are accepted and to be kept because they are related to the test objectives and can measure what the test intended to be measuring. These test items have acceptable degree of congruence with the test objectives. On the other hand, the test items with the IOC index of ≤ 0.5 should be revised or discarded as they have very low ability to measure what the test intended to measure.

Moreover, the experts strongly agreed that the content of the L-PESH Test reflects the objectives of the test. They believed that the L-PESH Test is appropriate to measure the English listening proficiency of the 4th year students majoring in the Tourism and Hotel Industry. The experts highly agreed that the content of the L-PESH Test covers various settings and situations found in service and hospitality routine work, and the specific language used in the L-PESH Test can be found in real working environments in the service and hospitality industry. They found that the quality of the recordings and the pictures are acceptable. However, they suggested that the speed of some talking in the recording should be slower and the pauses between test items should be longer. Some pictures that might be misleading should be discarded. Finally, they mentioned that the overall formats of the test and the time allotment is appropriate.

From the previous information, it can be concluded that the content validity of the L-PESH Test was satisfying, meaning that the test can measure what it claims to measure.

b) Criterion-related validity This usually includes any validity strategies that focus on the correlation of the test being validated with some well-respected outside measure(s) of the same objectives or specifications. For instance, if a group of testers was trying to develop a test for business English to be administered primarily in their institutions, they might decide to administer their new test and the TOEIC to a fairly large group of students and then calculate the degree of correlation between the two tests. If the correlation coefficient between the new test and the TOEIC turned out to be high, that would indicate that the new test was arranging the students along a continuum of proficiency levels very much like the TOEIC does - a result that could, in turn, be used to support the validity of the new test (Brown, 2000).

Criterion-related validity of this sort is sometimes called concurrent validity (because both tests are administered at about the same time). The criterion-related or concurrent validity of the new test in this study was estimated by having two tests, the L-PESH Test and the TOEIC, administered to the same group of test takers. Then the researcher calculated a correlation coefficient of the two sets of scores

(APPENDIX C) and determined the degree to which the scores on the two tests went together or overlapped. The correlation between L-PESH Test and the TOEIC was estimated and interpreted by applying Pearson product-moment correlation coefficient. The Pearson correlation coefficients of the L-PESH and TOEIC test scores was at 0.89, meaning that there was a positive high correlation between the L-PESH Test and the TOEIC. This would indicate that the new test, L-PESH Test, could arrange the students along a continuum of proficiency levels very much like the TOEIC did. This result finally supported the validity of the L-PESH Test.

c) Construct validity According to Brown (2000), in most cases, construct validity should be demonstrated from a number of perspectives. Hence, the more strategies used to demonstrate the validity of a test, the more confidence test users have in the construct validity of that test, but only if the evidence provided by those strategies is convincing.

In short, the construct validity of a test should be demonstrated by an accumulation of evidence. For example, taking the unified definition of construct validity, we could demonstrate it using content analysis, correlation coefficients, factor analysis, ANOVA studies demonstrating differences between differential groups or pre-test and post-test intervention studies, factor analysis, multi-trait/multi-method studies, etc. Naturally, doing all of the above would be a tremendous amount of work, so the amount of work a group of test developers is willing to put into demonstrating the construct validity of their test is directly related to the number of such demonstrations they can provide.

In this study, the construct validity was estimated by conducting a content analysis and an analysis of L-PESH Test Task Characteristics. Textbooks and teaching materials were analyzed to identify test construct. As this study focuses on assessing listening proficiency in English for the service and hospitality industry, the analysis put more emphasis on listening tasks to be used in test development. Consequently, the TLU in listening settings and tasks to be included in the test could be categorized into four main groups namely TLU listening tasks in Front Desk Service, Food and Beverage Service, Housekeeping Service, and Conference and Banqueting Service.

A number of books and materials used in teaching English for the service and hospitality in both private and public universities in Bangkok together with the majority of books related to the field available in university libraries (See list of books and materials in APPENDIX D) have been analyzed in order to provide/select TLU and TLU tasks for the development of "Listening Proficiency Tests in English for the Service and Hospitality Industry (LPESH Test)".

The procedures of how TLU situations are derived are as follows:

1. The researcher has conducted informal surveys and interviews with friends and lecturers in selected public and private universities about the curriculum and teaching materials used in English courses in the Department of Tourism and Hotel Industry in their universities.

2. After that, a library survey was conducted in order to finds more books used in teaching English for tourism and hotel industry.

3. The researcher went through each book, making a list of contents, including settings, situations, and language used.

4. Next, a very simple way of calculation, tallying, was applied to find the frequency of each setting and language used.

5. Finally, from this analysis, the overall test task characteristics in TLU situations have been found and categorized in Table 3.2 below:

Table 3.2 Test Task Characteristics to be Included in the L-PESH Test

The test task characteristics in TLU situations, presented in HIGHER frequency:	**The test task characteristics in TLU situations, presented in LOWER frequency:**
A. Front Desk Service 1. Greetings/ saying goodbye/ pleasing and thanking/ apologizing	**A. Front Desk Service** 1. Offering help and advice 2. Enquiries and giving information on recreation facilities

2. Taking hotel reservations: - enquiries and reservations - telephone enquiries: taking incoming hotel calls - asking for clarification - reception- reservations by phone - dealing with phone requests about hotel facilities and services - spelling on the phone, giving and understanding spelling - confirming reservations - changes to reservations - suggesting alternatives - giving information on hotel location, facilities, and prices - apologizing as in turning down reservations 3. Reception work: - receiving guests and making arrangements - checking in and checking out - asking for information and clarification - understanding customers' opinion and wishes 4. Paying bills: - payment enquiries - methods of payment - asking for the bill - explaining the bill - exchanging money 5. Making and dealing with complaints about the state of the room and slow or incomplete service. 6. Giving directions: indoor and outside. 7. Switchboard operator services.	3. Organizing excursions / tour operation: contact, negotiation, and local tour arrangement

B. Food and Beverage Service	**B. Food and Beverage Service**
1. Greetings/ saying goodbye/ pleasing and thanking/ apologizing 2. Taking restaurant reservations: - enquiries and reservations - telephone enquiries: taking incoming restaurant calls - giving information about restaurant - confirming restaurant reservation - saying time: opening/closing hours 2. Receiving guests and making arrangement: - presenting menus - recommending and describe dishes - taking orders for starters, main course, dessert, and drinks - room service: taking order on the phone e.g. a guest ordering breakfast 4. Paying bill: - payment enquiries - methods of payment - asking for the bill - explaining the bill 5. Making and dealing with complaints about food and drinks, service, and dining utensils	1. Giving instructions: mixing cocktails 2. Asking, comparing, and making recommendations: as in wine waiter taking order and describing wine 3. Explaining the use and purpose of dining utensils / equipment 4. Describing restaurant and kitchen 5. Menu planning 6. Purchasing and storage 7. Cost control and accounting
C. Housekeeping Service 1. Dealing with complaints on room facilities 2. Asking for / requesting for missing or extra room facilities, linens, towels, soaps, and so on	**C. Housekeeping Service** None

E. Career plan 1. Writing CV / resume and a letter of application 2. Attending job interview	**E. Career plan** 1. Describing jobs and workplace 2. Recruitment and job hunting 3. Stating one's attitudes towards a possible job

As this study focuses on assessing listening proficiency in English for the service and hospitality industry, the analysis put more emphasis on listening tasks. The researcher therefore selects the TLU listening situations and TLU listening tasks from the group of higher frequency only. Consequently, the TLU and tasks to be included in the test can be categorized into four main groups namely TLU listening tasks in Front Desk Service, Food and Beverage Service, Housekeeping Service, and Conference and Banqueting Service. Finally, in order to triangulate the findings of the content analysis, the researcher asked one expert, two teachers and two hoteliers to give comments on the L-PFSH Test task characteristics and the TLU tasks to be used in the test.

3.3.1.2 Reliability: In general, the test reliability is defined as the extent to which the results can be considered consistent or stable (Brown, 1996:192). Such consistency is desirable because the researcher does not want to base the decisions on an unreliable or inconsistent test that may lead to unreliable decisions or judgments. The degree, to which a test is consistent, or reliable, can be estimated by several methods.

In this study the researcher decided to use the Kuder-Richardson (KR 20) method to estimate the internal consistency of the test because it corresponds with the characteristics of the test. Moreover, KR-20 is good to be applied when test items were scored dichotomously (i.e., right or wrong). The test was given just once, and the scores from this test can be used to estimate test reliability. The received reliability value was at 0.91, meaning that the test had very high reliability. The test result would be consistent no matter how many times it was repeated.

3.3.1.3 Item Analysis: When a new test is developed, after a large number of test items have been written, one or more tryouts are to be conducted. The tryouts help the test writer select the best items, make improvements in weaker ones, and discard the very poor items.

Data analysis from the try-outs can identify weak or defective items; determine the difficulty of each item, its power of discrimination between good or poor students, and the effectiveness of distractors.

In this study, the researcher applied the item analysis program (classical model) initially developed by Chung Ten Fan to conduct the item analysis, and the expected value of item difficulty was set at 0.20-0.80, while the expected value of item discrimination power was set at ≥0.30.

The result of the item analysis from the pilot study was given in the following tables 3.3 and 3.4.

Table 3.3 Summary of L-PESH Test Statistics
(Pilot Version with 124 Test Items)

Total number of test takers =	47
Total number of test items =	124
High Group = 12	Low Group = 12
Mean = 79.94 Max = 109	Min = 50 Median = 81 S.D. = 14.70
Kuder-Richardson Reliability Statistics	
KR 20 = 0.91	SEM 20 = 4.49
KR 21 = 0.88	SEM 21 = 5.19
Good items that should be kept	83 items.
The items that should be revised	25 items.
The items that should be deleted	16 items.

From Table 3.3, it was found that the L-PESH Test (pilot version with 124 test items) had very high reliability (KR 20=.91) with 83 items to be kept, 25 items to be revised, and 16 items to be discarded. The highest score was 109 while the lowest

was 50. The mean score was at 79.94 while the median was at 81 with standard deviation of 14.70.

Table 3.4 L-PESH Test Summary (Pilot Version with 124 Test Items)

Type	Mean	Min	Median	Max	Std	Var
Test scores	79.94	50.00	81.00	109.00	14.70	216.10
Diff.index	0.65	0.11	0.66	1.00	0.23	0.05
Delta	9.73	0.00	10.60	18.60	4.77	22.78
Disc. Index	0.30	- 0.58	0.33	0.83	0.25	0.60
Biserial	0.33	- 0.46	0.34	1.27	0.29	0.08
Point-Biserial RPB	0.28	- 0.33	0.31	0.68	0.20	0.04

It was found from Table 3.4 that the mean of difficulty index of the test was 0.65 meaning that the test, in general, was not too difficult or too easy. As the expected value of difficulty index was set at 0.20-0.80, the level of difficulty index of this test is acceptable. Meanwhile, the mean of discrimination index was at 0.30 meaning that the test has good ability in discriminating good students and poor students. As the expected value of discrimination index was set at $\geq$0.30, the discrimination index of this test is acceptable as well.

The result from the item analysis was triangulated with the data gained from the students' attitude questionnaires. The following paragraph summarizes the results of student's attitude questionnaire from the pilot study.

From an analysis of 47 questionnaires returned, it was found that there were 7 male and 40 female test takers aged from 20 to 22. Twenty-four of them were from the Department of Tourist and Hotel Management, Bangkok University, and 23 of them were from Department of Career Sciences, Kasetsart University. Their average GPA was 2.75, and the average grades for English courses taken was B. Among these test takers, two of them have taken TOEIC.

The test takers agreed that the test format was satisfying and easy to follow though some pictures could not be seen clearly. They found that the test content was close to the content in English courses they have learned. The level of difficulty was

at average level for them. They did not find it too hard or too easy. The quality of the recording was good; however, the speed of the dialogues and speeches was a bit too fast. They further suggested that there should be longer pauses after each item for thinking and marking answers. They commented that there were too many test items in one test, and the test time was too long.

The following is about additional opinions about the test. They agreed with the idea of having a variety of English in the recording as they found it helpful and more realistic. They mentioned that in the real working situations they would not only encounter native speakers of English but also would encounter all varieties of English. They also mentioned that the institution should provide this kind of test for them before they graduate, and it would be better if all skills were included in the test. Regarding the strengths of the test, the test takers found that the test was useful and could help them improve their English before they enter the real job market. However, there were some drawbacks to consider. These included the speed and pause in the recording. The pictures were to be improved as well. The length of the test should also be considered as there were too many test items so it was too long. They found the test was useful to their studies and future career.

Conclusions and decisions could then be drawn from theses findings that the L-PESH Test (pilot version) should be revised in terms of its length, speed and pauses in the recording. Some pictures and test items that had very low statistical value should be revised or eliminated.

The test was then revised and put into the second draft including 80 test items which takes about 60 minutes to administer. The validity and reliability of the 80 item-test version was also calculated together with an item analysis. This time it was found that the value of test validity and reliability was higher than those of the first version (with 124 test items). The reliability coefficient (KR 20) of the revised version was at 0.89. This revised version of the test was used in the main study. See the revised version of the L-PESH Test (used in the main study) in APPENDIX B.

3.3.1.4 Summary of the L-PESH Test

The test is designed to measure listening proficiency in English for the service and hospitality industry of graduating students majoring in service and hospitality industry from private and public universities in Bangkok.

a) The purpose of the test: It is to be an ESP listening test to assess the level of listening ability of the graduating students majoring in the hotel industry.

b) The TLU situation and TLU tasks: They include three main hotel works in three departments: the front desk, food and beverage service, and house keeping. The tasks include face-to-face and telephone conversations in various situations.

c) Characteristics of the test takers: The test takers are 4th year students majoring in the service and hospitality industry, and have completed all required English courses in the curriculum. They are present students from four selected public and private universities in Bangkok.

d) Characteristics of the TLU situation: There are various situations concerning hotel work, such as at the front desk: checking in/out, paying bills, making complaints and dealing with complaints, taking phone calls, replying to inquiries, and so forth.

e) Characteristics of TLU task: There are more tasks on listening and speaking including face-to-face and phone conversations.

f) Definition of the construct to be measured:

Grammatical knowledge: simple statement, requests, apologizes, questions, and refusal.

Textual knowledge: Based on dialogs and talks related to job situations.

Functional knowledge: Functional English for hotel routine chores.

Sociolinguistic knowledge: Politeness, registers, Cultural

knowledge and the knowledge
of English as a global language.

Background knowledge: Similar background knowledge as they are students from the same major and educational basis.

g) Content of the test:

Organization of the test: The test is a one-hour, paper-and-pencil, multiple-choice test that consists of 80 questions divided into four parts:

Part I:	Photographs	10 questions (4-choices)
Part II:	Question-response	25 questions (4-choices)
Part III:	Short conversations	25 questions (4-choices)
Part IV:	Short talks	20 questions (4-choices)
Total		80 questions
Total time		1 hour

Time allocation: 1 hour for the whole test. Timing of questions is approximately 30 seconds per question. There is a thinking gap for about 10 seconds per question on the tape as well.

h) Scoring criteria

Criteria for correctness: As this is a multiple-choice test, each test item can have only one correct answer. It is an objective scoring, free from bias. The answer key is provided.

Scoring procedures: It can be either machine scored or manually scored.

i) Samples of the items: See samples of the test items in the test booklet in APPENDIX B.

j) Plan for evaluation the qualities of good testing practice

The test usefulness will be investigated in terms of reliability, validity, situational authenticity, international authenticity, impact, and practicality.

3.3.2 Students' attitude questionnaire, written in Thai.

The students' attitude questionnaire (APPENDIX E) was designed by applying the framework suggested in Dornyei (2002). The questionnaire was divided into three main parts:

Part One was to investigate general information about the test takers including gender, age, field of study, GPA, and the highest and lowest grades in English courses they obtained. Their experiences in taking other standard tests were also included in this part.

Part Two investigated the test takers' attitudes towards the test by applying five-point Likert scales. There were twelve items to be rated concerning major characteristics of the test such as test format, difficulty level, pictures, font, quality of the recording, content and test usefulness.

Part Three included three opened questions about the test strengths and weaknesses, a variety of English accents included in the recording, and whether there should be this kind of testing administered by their institutions before they graduated.

The validity of the student's attitude questionnaire was estimated by having three content specialists review it. The reliability coefficient of the attitudinal scale was investigated by applying Cronbach alpha. Its reliability coefficient was 0.72.

3.3.3 Test validating form (for the test reviewers)

The test validating form for test reviewers (APPENDIX A) was developed based on the suggestions given by Osterlind (1998) and Haladyna (1994). This form was divided into two main sections:

Section One: Construct Validation.

The reviewers were asked to read the item objectives and skills to be measured in each part of the test and consider carefully the degree to which the item was congruent, related to the skills. Then they rated the congruence according to this scheme:

H = high degree of congruence
M = medium degree of congruence
L = low degree of congruence or certainty

Additional reviewers' comments could be recorded in the space provided. After the first section was completed, the reviewers moved to Section Two: Content Validation.

Section Two: Content Validation.

In this section the reviewers were asked to answer each question concerning the test content with YES or NO only. Finally, descriptive statistics were applied to analyze the Index of Item-Objective Congruence (IOC) in order to estimate the construct and content validity of the test.

3.3.4 TOEIC, a standardized test to be administered by the TOEIC Thailand

The researcher contacted the TOEIC Thailand test center to administer the test with the corporate rate of 600 Baht per test taker. The test was administered at Kasetsart University. The test takers can use the test result to apply for a job in the real job market. The test result can be kept for two years.

3.3.5 Face - to - face and telephone interviews.

(See the list of guided questions for the interviews in APPENDIX F)

Face-to-face and telephone interviews were used in this study to get information on employers' needs on the level of language ability of their prospective employees, more details about the candidate's desired qualifications, the target language use in real situations, the process of recruitment, and the suggestions for universities to improve the students' ability in English.

In brief, there were two instruments used to gather data in the main study. The first one was the L-PESH Test, the revised version of 80 test items, and the test administration time was 60 minutes. The second one was the revised version of the student's attitude questionnaire.

3.4 Data collection

In the main study, there were 250 test takers purposively and randomly selected from four universities in Bangkok. Due to the test takers' time and transportation constraint, these 250 test takers took the L-PESH Test at their own institutions on the appointed dates and time. The test settings had been arranged in the similar test setting, meaning that the test was administered in a soundproof room with good quality of audio equipment. In order to avoid the problem of not having enough qualify headphones, the test takers listened to the test through loudspeakers instead of the headphones.

After the test administration had been done, the test takers were asked to fill out the student's attitude questionnaires within 15 minutes. The process of the test administrations in four universities took approximately two months in July and August, 2005.

3.5 Data Analysis

The test results and student' attitude questionnaires were analyzed as follows:

3.5.1 The L-PESH Test and its standard setting

The test results were analyzed by the Academic Testing Center of Chulalongkorn University applying the item analysis program (classical model) initiated by Chung Ten Fan. The findings (APPENDIX I) included test scores and test quality. Descriptive statistics were applied to analyze test scores for their means, median, minimal and maximal scores, standard deviation, difficulty index, and discrimination index.

The findings from the previous analysis were used in a standard setting process, establishing cut-off scores and their descriptors, in order to identify students' listening ability in English for the service and hospitality industry.

To establish the cut-off scores of the L-PESH Test, the mean score and the standard deviation of the score in normal distribution were calculated. Eight proficiency levels were established from these cut-off scores.

The L-PESH Test was a new test which was developed to measure listening ability in English for service and hospitality in which basic requirements include the use of a language in two main tasks; front of house and back of house tasks. Front of house tasks usually require a higher level of listening ability while back of house tasks require a lower level. Moreover, this test was not aimed at a high stake one at this stage. Three more skills of the test are needed so that standard setting can be set accurately. The researcher therefore decided to set the cut-off scores, based on framework suggested by Angoff (1971), Brown (1996), Morgan and Michaelides (2005), and Claycomb (1999). Finally, eight levels of listening ability could be established as Distinguished, Superior, Advanced-High, Advanced-Low, Intermediate-High, Intermediate-Low, Novice-High, and Novice-low.

In writing descriptors for each proficiency level, the researcher applied the information from an item analysis: difficulty index and a discrimination index together with information on each test item construct. The test items were ordered according to their difficulty and discrimination index. Then the test items were grouped into eight levels in order to define the construct (what ability the test takers were expected to perform if they got the answer in each item right). Finally, the description for each level could be defined.

The established cut-off scores and description were justified by five experts including 3 Heads of the Departments in Bangkok University, Kasem Bundit University, Kasetsart University, Rangsit University, and one HRD manager, one assistant manager from the Pan Pacific Hotel. To triangulate this justification, five test takers were drawn from eight levels of the cut-off scores to be interviewed and

their listening ability was checked by the teachers in each institution to confirm that the test takers really have the ability at the specified level. This process was conducted in four universities.

3.5.2 Hypothesis testing

To test the first hypothesis, stating that the new test, the L-PESH Test, can measure listening ability in English for the service and hospitality industry, Pearson Correlation Coefficient value of the TOEIC listening score and the L-PESH Test score was considered. The expected Pearson Correlation Coefficient value of the two sets of scores was set at >0.75. In this study, the Pearson Correlation Coefficient value was 0.89, meaning that there was a high positive correlation between the scores of the two tests. This means that the new test, the L-PESH Test can measure what it means to measure as well as the standard test does.

The second hypothesis, stating that the L-PESH Test can differentiate students into eight different ability levels of listening in English for the service and hospitality industry, was tested by looking at the ability descriptors of each performance level. According to the eight ability descriptors, it was found that the students who have the highest ability level, Distinguished, are able to perform tasks more than those whose ability levels are lower.

3.5.3 The student's attitude questionnaire

After the test administration at each university, the student's attitude questionnaires (revised version) were distributed to all test takers to be filled out within 15 minutes. Descriptive statistics was applied to analyze the data from these questionnaires.

3.6 Conclusion

To answer the first research question, "Can the L-PESH Test differentiate Thai graduating students' listening proficiency in English for the service and hospitality industry?", the Pearson correlation coefficients of the L-PESH Test and

the TOEIC were used to estimate the criterion-related validity of the new test, the L-PESH Test. The Pearson correlation coefficient of the two tests was 0.89, meaning that there was a positive high correlation between the L-PESH Test and the TOEIC. This indicated that the new test could measure the students' ability along the continuum of proficiency levels very much like the TOEIC did. This result supported that the L-PESH Test can differentiate graduating students' listening proficiency in English for the service and hospitality industry.

To answer the second research question, "What are the appropriate cut-off scores for each level of the listening ability?, evidence obtained from the test results together with the expert's opinions will be used to establish the cut-off scores based frameworks suggested by Angoff (1971), Brown (1996), Morgan and Michaelides (2005), and Claycomb (1999).

To answer the third research question, "What are the descriptors for each level of listening ability?", proficiency levels and descriptions suggested in the ACTFL and the TOEIC together with the information from the Can-Do guide of the L-PESH Test were applied to explain descriptors for each level of listening ability of the L-PESH Test.

To sum up, Chapter Three presents the research methods and the procedures of the study. The presentation covers five major areas including research procedures, subjects, research instrumentation, data collection, and data analysis. The research results and discussions will be presented in the next chapter.