CHAPTER III

RESEARCH METHODOLOGY

This procedure section includes discussion of: (1) research design, (2) population and samples, (3) the construction and validation of the research instruments, (4) data collection, and (5) data analysis.

I. Research design

Since the researcher looked at the effects of two independent variables that were the listening test formats and the use of English accent varieties as the listening test stimuli on one dependent variable which were the listening comprehension test scores, the study resulted in an experimental approach of research. The design was considered an experimental research that involved manipulation, control and randomization. The independent variables which were the use of the two different test response formats and the accent varieties of English were manipulated. This study followed the experimental design by employing the random selection and random assignment of groups to the manipulations.

The study was a *Mixed Randomized-Repeated Design* (Tabachnick and Fidell, 2001). That was, there were different subjects or cases at different levels of one randomized-group IV, and each case was measured repeatedly on one repeated-measures IV. This present study employed the simplest mixed design that was a two-way factorial with one IV of each type. The mixed factorial design was a combination of "within-subjects design" and "between-subjects design". It was a factorial design that included both between and within subject variables. This was a design in which all subjects were given the multiple choice and the short answer test format, and these two together served as a within subject factor. The participants were also divided into two groups. One group was treated with listening input of native speakers of English, while the other listened to the listening stimuli from nonnative speakers. The accent varieties of native and nonnative speakers served as a between subject variable.

The present research design consisted of one within subject variable that was the test format with two levels – multiple choice and short answer, and one between subject variable which was English varieties with two levels – native speakers and non-native speakers. The design can be also called a 2x2 mixed factorial design (Bates College Psychology Department, 2006) and can be illustrated as follows:

Table 3.1

2X2 Mixed Factorial Design

Varieties of English (Between-subjects)	***Test formats*** (Within-subjects)	
	Multiple choice (MC)	Short answer (SA)
Native Speakers	*Native MC*	*Native SA*
Non-native Speakers	*Non-native MC*	*Non-native SA*

This design is sometimes called a *split-plot* design (Field, 2005) due to its origin in an agricultural setting. A piece of land was split into several different plots or strips for treatment with different fertilizers or whatever treatments, and repeated measures were made along the length of each strip.

The justification for using this design is due to its power of repeated-measures analysis. In measuring the test format impact in each subject in two conditions (MC and SA) means that the subject is serving as his or her own control. The repeated-measures analysis controls for this. If the subjects vary a lot from one another, the repeated-measures analysis will have more power than ordinary two-way ANOVA (StatSoft, 2007).

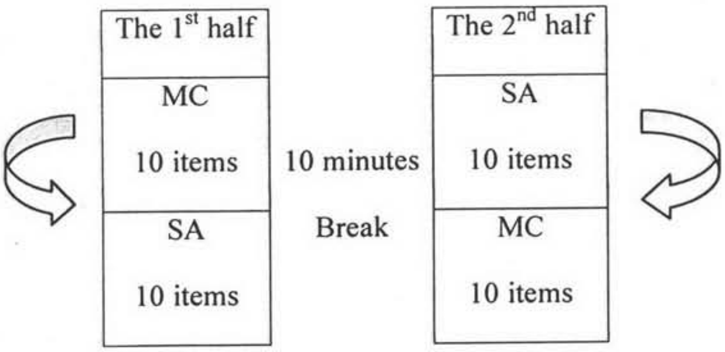
There are two major advantages to this type of design (Tabachnick and Fidell, 2001). The first is the test of the generalizability of the repeated-measures IV over the levels of the randomized-groups IV when the randomized-groups IV is a selected characteristic of cases. If a repeated-measures IV produces a different pattern of results at different levels of the randomized-groups IV, there is significant interaction and knowledge that the effects of the repeated-measures IV apply differently to different groups. The second advantage is the increased power due to the smaller error terms associated with the repeated-measure segment of the design. As mentioned by StatSoft (2007), this design has more power than an ordinary two-way

ANOVA. In repeated-measured analysis, differences among cases are assessed and removed from the repeated-measures error terms. The sum of squares for differences among cases is then partitioned into the randomized-groups effects and their error terms. These error terms are used with both the repeated-measures IV and the interaction with repeated measures, so power for these effects is increased.

In planning a kind of repeated-measures study like this present research study, it is important to consider the impact of repeated experience with levels of the IV and the DV on the performance of the subjects. The fact that the subjects might become more or less skilled in the performance of the test scores according to the order of the levels of IV is known as a carryover effect. If control is not exercised over the order in which formats of tests are taken, the carryover effect could be misinterpreted as an effect of that particular test format. For instance, if the subjects take the multiple choice part first, they might perform better on the short answer part that follows because they become more skillful. To control over the carryover effect in this situation, a counterbalancing technique was exercised. Since there were two test formats which were multiple choice (MC) and short answer (SA), the subjects were presented both formats in two possible orders that were MC and then SA, and SA and then MC. The order of the test formats presented to the subjects is described in the following figure:

Figure 3.1

The Order of the Test Formats



II. Population and sample

a. Population

The target population in this present study was the second-year English major students who enrolled in the Listening Comprehension Course in the second semester of the 2005 academic year at the University of the Thai Chamber of Commerce. The total population number was exactly 380. It could be expected that the students shared almost the same number of years studying English and all of them were in the same major of study so it was anticipated that they had positive attitude towards learning the language and they were interested in English before taking the course. The participants were homogeneous in terms of nationality and background knowledge as they were all Thai students in the same university. Most of them were about the same age and it could be assumed that they had similar culture, interest and educational background. There were more female students than male students. Female students accounted for approximately 80% of the total population.

b. Sampling Method

To fulfill the sufficiency and representativeness of the sample, the researcher then applied the method of stratified random sampling in which certain subgroups were selected for the sample in the same proportion.

The steps in the sampling process were as follows:

1) The target population was identified. The number of the second year students in the academic year 2005 who registered in the course was 380. There were 8 groups of the students in the day program and 3 groups in the evening program.
2) Using a table for determining a sample size from a given population (Krejcie and Morgan, 1970:608), with the reliability of 95% and error rate of 5%, the sample size required to be representative of the 380 students was 191. The researcher decided to have the sample size of 192.
3) The samples of 192 were randomly selected from the 11 groups. Using a table of random numbers (Fraenkel and Wallen, 2000:646-7), the student samples were chosen in the same proportion as they existed in the population.

4) The samples of 192 were divided in half. Using 'mechanical matching' (Fraenkel and Wallen, 2000:294), they were randomly assigned into group 1 and 2 equally. The subjects in group 1 then took version A test, while the group 2 subjects undertook version B test.

The reason that the researcher employed stratified random sampling was because of the fact that the students were assigned into groups according to their study performance levels from the previous English courses; for example, the students in group 4 performed better than the students in group 5, etc. Consequently, a stratified random sampling technique virtually ensured that the student of all performance levels would be selected for the sample in the same proportions as they existed in the population.

However, in an attempt to increase the likelihood that the two subject groups would be equivalent, pairs of 192 subjects chosen were matched on their final listening comprehension test scores. It was possible to obtain the subjects' listening comprehension scores from the final test because the experiment was conducted after the final examination in February 2006. Mechanical matching was a process of pairing two persons whose scores on a particular test were similar. The 192 subjects in this study were pair-matched using their final listening comprehension test scores. The subjects who had similar scores were matched and assigned into groups 1 and 2. To prevent eliminating some subjects from the study because there were no 'matchees' for them, the scores that were +1 or –1 were accepted to be equal in this situation. For instance, a subject who obtained 15 points was pair- matched with a subject who obtained 16 points. After the matching was completed, a check for the equivalence of the two groups was made. The means of the two groups were almost identical in that group 1's mean was 20.708, S.D = 5.608, and group 2's mean score was 20.739, S.D. = 5.601. However, to ensure that the two groups were not statistically different, the two groups' means were compared using the Independent-samples *t*-test. The *p* value from the *t*-test analysis showed that the two groups were statically equivalent ($p = .969$). This confirms that group 1 and group 2 were statistically comparable and equal. Figure 3.2 illustrates the sampling steps in the present research study and Table 3.2 and Table 3.3 show the result of the *t*-test for the two groups. The samples' scores in mechanical matching are presented in detail in Appendix E.

Figure 3.2

Sampling Technique

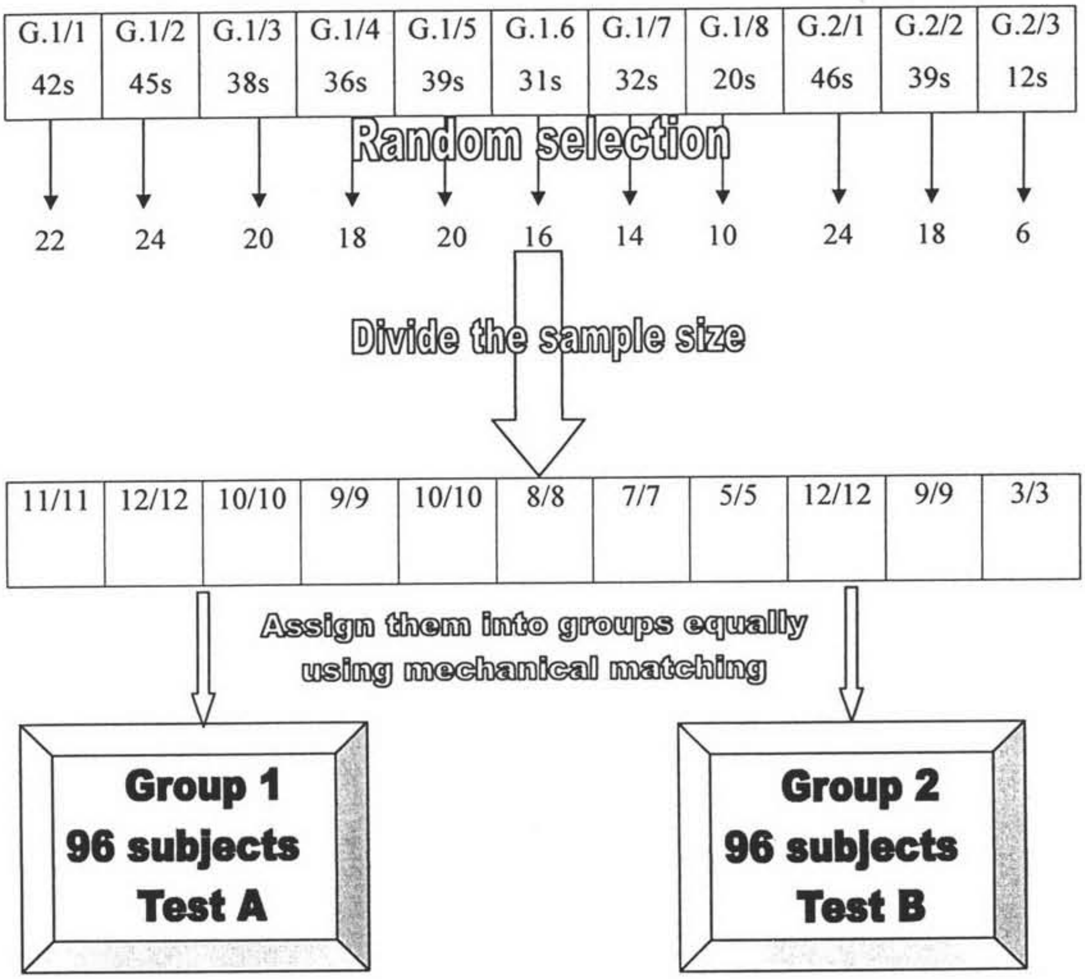


Table 3.2
Mean Scores of the Two Groups

	group	N	Mean	Std. Deviation	Std. Error Mean
Listening scores	1.00	96	20.7083	5.60811	.57238
	2.00	96	20.7396	5.60121	.57167

Table 3.3
Independent Samples Test for the two Groups

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Listening Scores	Equal variances assumed	.006	.939	-.039	190	**.969**	-.03125	.80896	-1.62695	1.56445
	Equal variances not assumed			-.039	190.000	.969	-.03125	.80896	-1.62695	1.56445

III. Research instruments

Two kinds of instruments were used in this study: listening comprehension test version A and B, and retrospective semi-structured interview questions used with a modified matched-guise technique to investigate the subjects' attitudes and preferences towards varieties of English.

a. **Listening Comprehension Test**

The listening comprehension test is described into 2 parts: test construction and test validation. The test construction section describes steps in developing the tests. These include structure of the test, test tasks, test version and selection of speakers' countries. The test validation part describes steps and statistics used in validating the test.

Test Construction

1) Test format selection

The test format variables were chosen to be studied to see whether varying the formats would result in any differences in test takers' ability. There were many test response formats other than the multiple choice that were used in the Listening Comprehension course However, the four commonly used formats in the classroom at UTCC were multiple choice, true-false, gap-filling and short answer. In order to be certain about the selection of the appropriate test formats to be studied, the pilot study was conducted in February 2006 with 99 students. These 99 subjects were selected randomly from the same population of the main study. The pilot test contained multiple choices, true-false, gap-filling and short answer format. The four formats were chosen because of their preference in the course and the students were exposed to these test formats in their class.

The null hypothesis was set as *there was no significant difference in the scores obtained from the four different test response formats*. To test the hypothesis, a One-Way Within ANOVA was conducted to examine the differences among the scores. The summary table of the repeated measures effects in the ANOVA with corrected *F*-values is below.

Table 3.4
Test of Within-Subjects Effects for the Four Formats

Source		Type III Sum of Squares	df	Mean Square	F	Eta Squared
FORMAT	**Sphericity Assumed**	**1390.634**	**3**	**463.545**	**77.693***	**.442**
Error(FORMAT)	Sphericity Assumed	1754.116	294	5.966		

* $p \leq .05$ (two-tailed).

The table illustrates that there was a significant within subject effect for the test formats, F (3, 294) = 77.69, $p < .05$, with an effect size of $\eta^2 = .442$. This suggests that test takers' listening ability was affected by the test formats used. The effect size was large (Hopkins, 2002). The Eta squared value was .442 which means that the test format factor by itself accounted for 44 % of the overall (effect+error) variance (UCLA, 2006). Another way to interpret effect size is to compare them to the effect sizes of differences that are familiar. The value of $\eta^2 = .442$ was transformed to Cohen's d of 1.78. Cohen describes an effect of more than 0.8 as 'grossly perceptible and therefore large' and the size is large enough to be easily visible (Coe, 2000).

After the null hypothesis was rejected, a multiple comparison was conducted to find out where the differences lied among the means. Bonferroni was selected as it is a very conservative test and it is powerful for a small number of pairs (Tabachnick and Fidell, 2001). The results from a multiple comparison are reported in the following table.

Table 3.5
Multiple Comparison Test for the Four Formats

(1) FORMAT	(2) FORMAT	Mean Difference (1-2)
MC	GF	-5.020*
	TF	-3.949*
	SA	-2.697*
GF	MC	5.020*
	TF	1.071
	SA	2.323*
TF	MC	3.949*
	GF	-1.071
	SA	1.253*
SA	MC	2.697*
	GF	-2.323*
	TF	-1.253*

* $p \leq 0.05$ (two-tailed)

The results obtained from a multiple comparison show that:

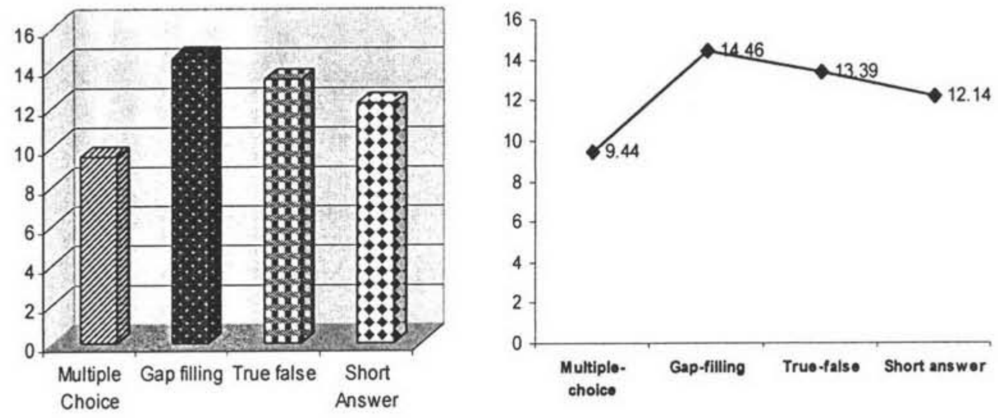
1. *Multiple choice test format* scores are different from gap-filling, true-false and short answer test format scores;
2. *Short answer test format* scores are different from multiple choice, gap-filling and true-false test format scores;
3. *Gap-filling test format* scores are not different from true-false test format scores.

The graphs in Figure 3.3 display means of the four test formats.

The scores show that the test takers in the pilot study found the gap-filling test quite easy. The most obvious problem with this situation was that the test takers filled in the blanks without trying to understand the passage at all. They treated the passage as a normal cloze test, and filled in the blanks with the words they had heard. They did not really listen to the passage for comprehension. For this reason, the gap-filling format, in this situation, was not truly a listening comprehension test. Furthermore, a multiple comparison analysis reported that the scores from the gap-filling and the true-false format were not statistical different. The students also found the true-false test as easy as the gap-filling test format, although the mean score from the true-false test showed that it was a little harder than the gap-filling test. Because gap-filling and true-false formats were found to be too easy and the researcher was not certain whether the gap-filling format really measured comprehension compiled with the fact that the true-false allowed too much guessing; the researcher selected the multiple choice and short answer format to be investigated further for the main study

Figure 3.3

Mean Plots of the Four Test Formats



2) Background of the Test

The Listening Comprehension course, for which the test was designed, made use of the current approach of teaching methods – eclectic methods in order to help the students to arrive at the meaning of language input effectively. The instructional design of the course was based on the following behavioral objectives:

The students should be able to:

1. Guess the topic and identify main ideas;
2. Get details from the extracts;
3. Understand and complete the paragraphs;
4. Get the meanings of the vocabularies used in context and draw correct conclusions and valid inferences about social situation, the speaker's intent or the general context.

The concept of communicative teaching approach, that the language used for the purpose of communication, in a particular situation and for a particular purpose, has led the move towards communicative testing. Buck (2001) believed that the communicative testing movement is still a very positive influence in language testing. Therefore, this listening test was developed based on the communicative approach and the language samples only took place in authentic context and for communicative purposes (see Appendix L for the course description).

3) Purpose of the test

The listening test designed was an achievement test. It was based on criteria and the objectives were targeted for the class. It was a low-stakes test for the purpose of providing evidence of the test takers' ability to participate in listening tasks given in the course units. The purpose of this test was to measure students' control of specific listening sub-skills, lexical and grammatical forms used to perform a particular function. Interpretations of scores were used as a basis for assigning course grades. Also, the degree to which students meet minimum standard of mastery of the content of instructional units was considered from the results of the test. Decisions about instruction included determining what parts of each unit have been effectively learned and what parts might need revision were drawn from the test results.

4) The questions used to evaluate comprehension ability

The test was designed to evaluate the students' mastery of the content of the teaching units. The questions were developed to measure the students' ability to understand the following:

1. **The context**: Who are the speakers? Where are they? What are they doing? What is the relationship between the speakers?
2. **The motivation of the speakers**: Why does one speaker say something? What does the speaker want to do? Why does the speaker do something?
3. **The key words that the speakers use**: What expression does the speaker use?
4. **Main ideas**: What is happening in this conversation?
5. **Paraphrases of meaning**: What does the speaker mean by saying something? What do we learn about the speaker?
6. **Inferences about the meaning**: What is the topic of this conversation? How does the speaker feel about something? What do you think a speaker will do?

Questions varied in difficulty. This listening test questions include:

- Verbatim questions – questions that require students to remember specific words.
- Synthetic questions – questions that require students to piece together information and paraphrase ideas.
- Analytic questions - questions that require students to analyze the meaning and draw inferences.

(Rost, 2002:183-186)

A mixture of questions was used in this test. The test required that the test takers use short-term memory well. They need to first listen to and understand a conversation extract, and then listen to questions about the conversation and select written answers or write short answers by themselves. A mixture of phrase types was used in this test. This listening technique of using whole phonological phrases reinforces the idea that listening means hearing complete phrases and pause units rather than listening for individual words and sentences.

The test takers were given two supporting conditions concerning comprehension questions and taking notes while listening:

1. The test takers were allowed to preview questions when they listened to long texts and were supposed to answer more than 2 comprehension questions from one dialogue input.
2. The test takers were allowed to take notes if they felt the need to do so but it was not a requirement for the test situation.

5) Source of text

In selecting appropriate texts for the test, the test writer was concerned with the test conditions under which listening activities should be carried out. Special attention was paid to topic familiarity, language difficulty, length of the text, text structure, and the strategy to be tested by the text.

Texts were initially selected from two main course books that were used in the last two academic years, 2003 and 2004. These books were Impact Listening 3 (Harsch and Wolfe-Quintero, 2002) and Listen In 2 (Nunan, 2003). Also in the classroom, the students had exposure to other texts from various sources such as TOEFL and TOEIC tests, radio announcements, news, etc. All the texts used in class were commercially made and mainly spoken by native speakers of English.

Research has indicated that in selecting texts for examination it is the degree of students' familiarity with the topic that has a major effect on their performance (Weir et al., 2000). Topic familiarity emerged as a powerful factor at all levels of test takers' proficiency in listening comprehension tests (Schmidt-Rinehart, 1994).

A crucial part of the dialogue selection was to ensure that the students were reasonably familiar with the topics of each of the texts that had been used in the course. However, it was almost impossible for the students to remember any of the text as they had listened to hundreds of speech sample inputs during the semesters. Furthermore, all the sample listening texts were opened for the students in class only once so it was likely that they would not recognize the dialogues chosen for the present study instruments. The difficulty level of language in the two textbooks and other sources used in the classroom was very similar as they had been selected and used as the main course materials for the two different academic years. The contents and topics in the two books were quite the same, however, only the topics that were similar in the two books were chosen to appear in the test. The dialogue selected to be included in the test are displayed in Appendix G and Appendix I, and will be discussed further in the test validation section.

6) Test Specifications

The objectives of the Listening Comprehension course together with all the course materials used in the course were gathered and studied to form the test specification. The specification was based on the objectives and contents of the course. The details of test specification together with text script and the test are presented in Appendix A, G, H and I. Based on the test specification developed, 100 items questions were constructed and later these draft items were tried out with the 32 students for the first time in May 2005. After that 80 good items were selected for the pilot study later in February 2006. Table 3.6 presents the objectives and numbers of the test item

Table 3.6

Objectives and Numbers of Test Items

Objectives	Item	Number of Items for each format			
		MC	GF	TF	SA
1. Guess the topic and identify main ideas	5,50,57,61,62	1		2	2
2. Get details from the extracts	1,2,6,8,12,13,14,15,16,20, 42,43,44,45,46,47,48,49,51 52,54,55,58,59,60,65,66,67 68,70,71,72,73,74,75,76,77 78,79,80	12		15	13
3. Understand and complete the paragraphs	21,22,23,24,25,26,27,28,29 30,31,32,33,34,35,36,37,38 39,40		20		
4. Get the meanings of the words used in context and draw correct conclusions and valid inferences about social situation, the speaker's intent or the general context	3,4,7,9,10,11,17,18,19,41,53,56,63,64,69	7		3	5

7) Test structure

The listening comprehension test designed was an achievement test. It was based on the criteria and the objectives that had been targeted for the class. It was a low-stakes test for the purpose of providing evidence of the test takers' ability to participate in listening tasks given in the course units. The purpose of this test was to measure students' control of specific listening sub-skills, lexical and grammatical forms used to perform a particular function. The test had been developed and validated in May 2005. In the first trial, the test contained 100 items. After the first trial, only good items were selected to be included in the original version of the test. At this stage, the test contained 80 items designed with four response formats – multiple choice, gap-filling, true-false and short answer. Finally, the test were improved and validated again with the 99 samples drawn from the population of the present study. The final test version used in the main study contained 40 items. There were 2 sections with the two test formats (MC and SA) and each part consisted of 20 items. The test designed lasted exactly 30 minutes. The following table shows the design of the test.

Table 3.7

Listening Comprehension Test Design

Part	Format	No. of Items	Time (minutes)	Task descriptions	Text difficulty*
1.	Multiple Choice (short text)	10	7½	Test-takers listen to monologue or interactions, about 30 sec. long and answer one or two questions by choosing the answer from four choices	2.942**
2.	Short Answer (short text)	10	7½	Test-takers listen to monologue or conversations, about 30 sec. long and give short answers to the one or two comprehension questions	2.848**
3.	Short Answer (long text)	10	7½	Test-takers listen to long conversation about 2 minutes long and give short answers to 5 questions.	2.068**
4.	Multiple Choice (long text)	10	7½	Test-takers listen to long conversations about 2 minutes long and answer 5 multiple choices questions.	2.568**

** Text difficulty refers to the average scores from teachers and students' opinion for the conversation used in the test.*

***The average scores from the 4-point scale: very difficult =1, difficult = 2, easy =3, very easy = 4.*

From Table 3.7, the test designed could control the following:

1. the length of the text spoken;
2. the order of the test formats;
3. the number of test items for each format;
4. the time spent for each format;

As for the difficulty of the text used in the test, the researcher tried to control the difficulty of each part to have the closest difficulty level. On average, the text should not be too easy or difficult for the test takers. The scores from teachers and students' opinion should not be below 2 and higher than 3. Any dialogue used that was reported to be too difficult or too easy would be eliminated at the test pilot stage.

8) Scoring Method

Criteria for correctness: The answers for all items in the multiple choice section were an objective type; one correct answer received one point credit. For parts 2 and 3, which required the test-takers to give short answers of one or not more than three words, certain possible answers were expected. The test takers needed to give answers that corresponded to the answer key; one point would be given to the answer that touched the key set. In this part, correct spelling was not required as long as the answers given by the test takers were comprehensible.

Procedures for scoring the response: The test was scored according to a scoring key developed on the basis of the test writer's script. There was only one test marker who, in this case, was also the test writer.

Explicitness of criteria and procedures: the test takers were informed in general terms about the scoring criteria stated in the test paper.

9) Test versions

Since the study aims to compare the effect of English varieties used in the listening comprehension tests, the original tests designed and validated were transformed to form two versions of the test – test A and test B. Test A, the native speaker version, used only the voice of native speakers of English as the listening test stimuli, while test B, the nonnative speaker version, used the voice from nonnative speakers of English as the input. The content and construct of the two tests were kept strictly in the same manner as the original test version. Only the input stimuli were spoken by the speakers of different English varieties.

10) Selection of speakers' countries

The study required the test takers to encounter a range of English varieties; however, the experimental design limited the number of accents that could be tested. Given these limitations, three countries of the English native speakers were chosen for Test A and four countries of the nonnative speakers of English were selected for Test B. These countries were chosen according to statistics reported by the two government offices – the Board of Investment (BOI) and the Tourism Authority of Thailand (TAT) through their websites in 2006 (BOI, 2006; TAT, 2006). The goal of the study is to examine the influence of English spoken by people from different parts of the world on listening comprehension. Therefore, the people from the countries that the test takers who are Thai are likely to encounter were selected according to the amount of their investment in Thailand and their arrivals to the country.

Table 3.8

Major Foreign Investment in Thailand

	2004	**2005**	**2006**	**2004**	**2005**	**2006**
Countries	**No. of Projects Approved**			**Total Investment (Million Baht)**		
Japan	350	354	353	125,932	171,796	115,200
Europe	81	131	118	32,980	48,012	21,174
USA	37	48	46	30,397	8,689	71,407
Singapore	74	69	62	18,239	14,422	18,750
Taiwan	53	57	63	10,607	16,456	10,472
Hong Kong	23	18	18	14,317	2,222	10,031

(BOI, 2006)

Table 3.9

The International Tourist Arrivals to Thailand by Nationality

Country of Nationality	**% share 2005**	**% share 2006**
East Asia	**55.55**	**55.15**
Malaysia	11.93	11.51
Japan	10.39	9.49
China + HK	9.12	9.59
Singapore	5.65	4.97
Europe	**24.74**	**25.26**
United Kingdom	6.72	6.15
America	**7.24**	**6.68**
USA	5.55	5.02
South Asia	**4.71**	**4.57**
India	3.31	3.33
Oceania	**4.48**	**4.71**
Australia	3.72	3.98

(TAT, 2006)

The statistic data from the tables indicates that the major foreign investors are Japanese, European, Taiwanese, American, Hong Kongian, and Singaporean. This information agrees with the number of the foreign arrivals to Thailand reported by Immigration Bureau. Three countries with English as their native language were reported to visit Thailand the most. They were the USA, United Kingdom and Australia. Consequently, people from these countries were chosen for Test A. For Test B, people from Japan, Malaysia, China were included in the test as they were the major investors and accounted for the highest number of tourists. However, considering the amount of the investment in Thailand, Singaporeans has invested a large sum of money in the past three years. They are second to Japanese in terms of investment amount in Thailand. Therefore, the Singaporean variety was also included in Test B.

Test A: Native speaker varieties version (NS)

1. United States of America
2. United Kingdom
3. Australia

Test B: Non-native speaker varieties version (NNS)

1. Japan
2. Malaysia
3. China+Hong Kong
4. Singapore

In selecting the speakers' voice to be recorded for the test, the voice quality was carefully controlled. The basic requirements that were set out to control the voice variation were:

The speakers selected had to be:

1. educated – a degree holder;
2. adults – between 25 to 50 years old;
3. fluent in English for nonnative speakers.

There were both males and females for each accent. The quality of the voice recorded was very clear because it was done in the sound studio by an expert technician. These speakers were told to speak in their natural style and they were told that their voice would be used for a listening test. Therefore, they tried to speak as clearly as they could.

11) Test Validation

In the previous section, the test development procedure was described. For the validation of the test in this section, the first step was to have a test trial on a small but, representative sample of the potential test population and then the second step was to use statistical analysis to verify the test.

Listening Comprehension test trial

The test was first trialed in May 2005. The second –year students who had completed the Listening Comprehension Course in the second semester of 2004 academic year were invited to do the test in the listening laboratory at the University of the Thai Chamber of Commerce. The researcher had asked about 50 students from 8 different groups to attend and 32 students came and took the test. These students represented a wide range of ability. They were students who gained high, middle and low scores from the course. However, they were willing to participate in the test trial and completed the test seriously just like they were taking their final examination. They were allowed to read the questions before completing the test. Moreover, they were given the questionnaire to assess the topic familiarity and language difficulty of each conversation (see Appendix C). Therefore, after listening for each conversation and completing the answers, the tape was stopped for about 30 seconds in order to give the students time to complete the questionnaire.

Further, the questionnaire sheets were also sent to five teachers who taught in the Listening Comprehension course to ask for their opinions on (1) suitability of conversations selected for the test; and (2) the comprehension questions asked in the test. This was the comparison among the four test formats. The results of questionnaire survey from the teachers and the students are displayed in Appendix D.

The results of the questionnaire analysis were satisfactory. The degree of familiarity with each passage, the language level of the passages all fell within the desired ranges and extremes had been successfully avoided. Considering the data obtained from the teachers' and students' opinions together with the result of the first trial, the test was revised and improved then administered again with the 99 participants in the pilot study to validate the test in February 2006. In the pilot stage, the test contained 80 items with 4 test formats. At this stage, according to the result of the pilot study mentioned earlier, the researcher decided to use only multiple choice and short answer formats for the main study. Therefore, only the data from the two test formats were analyzed in the further statistical validation process.

The test data from multiple choice and short answer test formats were entered onto a computer and analyzed using SPSS. Descriptive statistics were generated for the test at the item, section and whole test. Firstly, the total scores for the test obtained by the test takers were investigated as follows:

Table 3.10

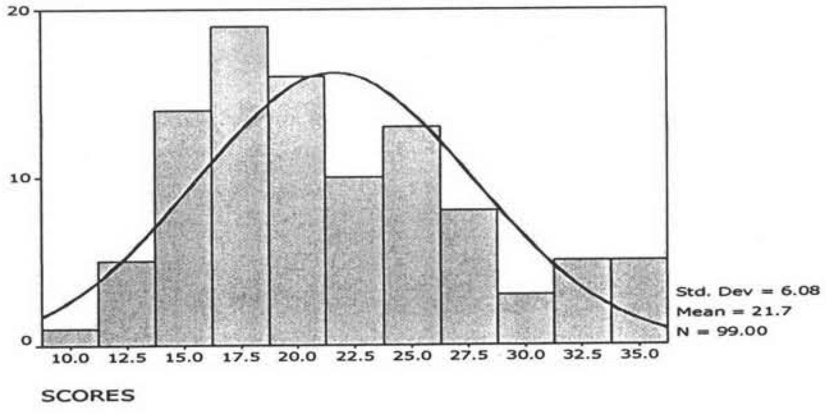
Descriptive Statistics of the Pilot Test Version

Statistics		**Values**	**Std. Error**
Mean		21.6566	.61120
95% Confidence Interval for Mean	Lower Bound	20.4437	
	Upper Bound	22.8695	
5% Trimmed Mean		21.4630	
Median		21.0000	
Mode		18	
Variance		36.983	
Std. Deviation		6.08136	
Minimum		10.00	
Maximum		35.00	
Range		25.00	
Interquartile Range		9.0000	
Skewness		.509	.243
Kurtosis		-.554	.481

There were 99 students who took the test. The highest score obtained was 35 out of 40 and the lowest score was 10. The difference between the highest and the lowest score in the pilot test was 25. The score which was the center of the distribution was 21. The average score was 21.64 which was quite close to the center of distribution. The most frequently obtained score was 18. Since the mean, the mode, and the median were not the same in this analysis, the distribution of scores was not normal. Furthermore, the skewness value indicates that the shape of distribution was not normal as it was not close to zero - 0.509. The histogram of the total scores in Figure 3.4 illustrates the distribution of scores for the pilot study.

Figure 3.4

Score Distribution for the Pilot Test



The standard deviation (SD) was 6.08. The larger the SD, the more variability from the central point in the distribution; if we consider the mean at 21.65, the SD of 6.08 was quite large. The scores for the pilot test version were widely spread so it is interpreted that the participants' ability in taking the listening comprehension test was not similar and they were a heterogeneous group.

Further, the test was analyzed by the Classical Test Item Analysis in CTG Package version 8 (Sukamolson, 2004) to validate the test concerning qualities of choices, reliability and difficulty. The statistical results show that the difficulty level of the test is about right for the participants (delta = 12.440). Suggested delta value should fall between 9.5 and 16.5 (Sukamolson, 1999) so the test is not too easy or too difficult. The reliability value of the test is good (KR-20 = 0.811) since a value that is close to 1.00 is desirable and the minimum value of 0.75 for KR-20 is acceptable to show that a test is reliable (Tulane University, 2006). The test statistics are summarized in the Table3.11, and their details are included in Appendix K and Appendix L.

Table 3.11

Test Statistics Summary

	Mean	**Min**	**Median**	**Max**	**SD**
Test scores	21.657	10.000	22.500	35.000	6.081
Difficulty index	0.541	0.071	0.515	0.960	0.511
Delta	12.440	5.976	12.449	18.921	3.014
Discrimination index	0.374	-0.038	0.288	0.615	0.186
Biserial (RBIS)	0.455	0.017	0.391	0.766	0.164
Point-Biserial (RPB)	0.341	0.012	0.285	0.557	0.129
Kuder-Richardson Reliability (KR 20)	.811				

b. Semi-Structured Interviewing

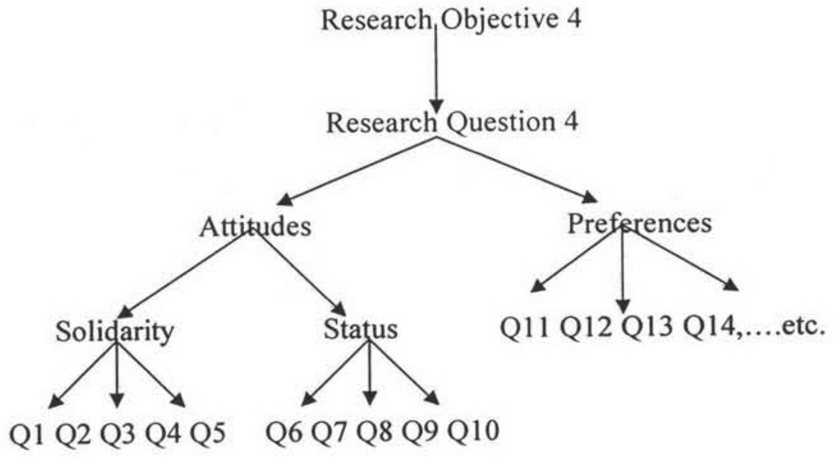
1. The Retrospective Semi-Structured Interview Questions

To investigate the attitudes of the subjects towards varieties and their preferences towards using English varieties as the listening input, the semi-structure interviewing was used. This type of interview involves the implementation of a number of predetermined questions and special topics. These questions were typically asked to each interviewee in a systematic and consistent order, but the interviewer was allowed freedom to probe beyond the answers to the prepared structured questions. The present study uses retrospective inspection to increase specificity. Here, the interviewees were supported in recalling a specific situation by playing some part of the conversations they had heard for the listening comprehension test, then they responded to the interview questions.

The term 'semi-structured' suggests a certain degree of standardization of interview questions and a certain degree of openness of response by the interviewer. All interview questions were structured with the purpose of the research question 4 set to investigate attitudes and preferences of the test takers. Figure 7 illustrates how interview questions were structured.

Figure 3.5

Structure Model of Interview Questions



The selected 30 subjects were recalled by listening to short speech samples of each variety, and they corresponded to the interview questions concerning attitudes and preferences. The present study used the traits studied by Hiraga in 2005 to investigate the participants' language attitudes in the questionnaire. Hiraga's choices of adjectives were chosen not only because her study is the latest in the field but also she had extensively revised all adjectives proposed to evaluate language attitudes in the preceding research studies and concluded with a list of ten very concise adjectives to study 'status' and 'solidarity' traits. Hiraga (2005) employed Factor Analysis to verify that the various adjective words were clearly divided into two response dimensions as shown in Table 3.12.

The traits chosen for investigating the solidarity dimension were sociable, sincere, comforting, friendly, reliable, and the traits for investigating the status dimension were educated, intelligent, wealthy, successful and elegant. The interview questions containing the 10 adjectives were asked and the subjects, after listening to the speech, responded by indicating whether these adjectives applied to the speakers by rating their opinion on the 4-point scale of agreement (see Appendix J).

Table 3.12

The Rotated Component Matrix of Trait Factors

	Component 1---	Component 2
sociable	0.16100	**0.66700**
educated	**0.91700**	0.07121
sincere	0.01569	**0.68800**
intelligent	**0.84900**	0.16700
comforting	0.14600	**0.79300**
wealthy	**0.92100**	-0.01766
friendly	-0.07141	**0.82100**
successful	**0.81500**	0.10700
elegant	**0.89700**	0.09239
reliable	0.40600	**0.48100**

(Hiraga, 2005:294)

2. The Modified Matched-Guise Method

For most language attitude studies, the matched-guise technique is the most frequently used (Hiraga, 2005). The major principle of the matched-guise technique is to examine only actual language varieties and to avoid control of other variables such as the voice quality of speakers, the content of texts, or the personality of speakers in the experiment situation. This technique requires that the passages be read by the same speaker who can pronounce all varieties correctly. However, the present study with its emphasis on subjects' reactions to target language accents suggests the use of a modified matched-guise technique, the 'verbal guise' method (Dalton-Puffer et al., 1997). Instead of one speaker assuming different guises, several speakers were used on the stimulus tape. In the present study context, it is practically impossible to find speakers who are equally convincing in several guises. This means that variables like voice quality could be controlled only minimally.

There were 14 speakers for the seven varieties. This means there were two people who represented each variety. The purpose of having two people from each variety was to check whether the respondents had consistency in responding to the same accent. The subjects were given instruction without identifying which varieties were included. The speakers were asked to read a short text on the same topic which was emotionally neutral and which also tied in with the university setting of the study. The participants were told that the test was done in the interest of finding the most appropriate English teacher. In reality, people react to speech in specific situations and the same voice or speaker may well get different evaluations in different contexts (Giles, 1992). It was more than likely that the subjects would construct a context for themselves if a specific situation was not provided, and this could lead to misinterpretation of the subjects' attitudes evaluation. The selected 30 subjects listened to 14 speakers who spoke the same dialogue:

"I help students pass university entrance exams. I sometimes worry about them and their futures because they don't know what they want to study in college, or what kind of job they want in the future. A lot of my students go to college because their families expect them to. Many of them think that once they pass the entrance exam their future is guaranteed. That's a mistake. I tell them, 'Passing an entrance exam is just the beginning. To find a satisfying career you have to be able to answer the following questions: What do you want to learn about? What lifestyle do you want? What are your goals?"

The stimuli were presented one at a time because previous work indicated that ratings of accentedness might become slightly harsher with repeated hearing of an utterance (Munro et al., 2006). Upon hearing an utterance, each participant was instructed to respond to the 10 questions starting with 'do you think this person is sociable?' by giving the answer in rating scales (Appendix J).

After these ten adjectives questions, another seven questions concerning their preferences when varieties of English existed as the test stimuli were followed. The questions asked were:

1. *Are you able to recognize different varieties of English?*
2. *Which varieties do you find easy or difficult to comprehend?*
3. *Do you find different accents equally pleasing?*
4. *Do your judgments depend on the voice of the speakers or the content of the utterance?*
5. *Do you like the listening comprehension test to incorporate varieties of English?*
6. *Does the inclusion of English varieties make you uncomfortable?*
7. *Does the inclusion of English varieties make the listening test more difficult or easier?*

The content validity of the interview questions were validated by three experts. The experts consist of one native speaker who holds a master's degree in teaching English and has taught listening and speaking English for more than 15 years. The other two content specialists hold a doctoral degree in applied linguistics and sociolinguistics. The three experts found the interview questions acceptable and valid.

IV. Data collection

The study aims to inspect the effect of using English varieties as the listening stimuli and the effect of test format variation on test takers' listening comprehension test scores. Consequently, the experiment consisted of several steps.

Data collection for the pilot study

1. Initially, the test was tried out with small groups of 32 students in May 2005. This group of students had studied in the Listening Comprehension Course from November, 2004 to March, 2005. The

teachers and students were also given the feedback questionnaire to ask their opinion and perception about the content and format of the test. The purpose of this stage was to validate the content of the test.

2. Then in February 2006, the sampling process was conducted at the end of the semester. The estimated sample size was 192. The students who were not selected for the main study were included in the second pilot study.
3. After the first listening comprehension test was revised and improved, it was tried out again in the pilot stage. Piloting the listening comprehension test for the second time happened after the Listening Comprehension Course was taught for about 16 weeks. In order to be certain that all the course content was taught to the participants, the pilot versions of the listening comprehension test was administered after the final week of instruction. 99 students participated in the second pilot stage. The purpose of this stage was to find the reliability value and difficulty index of each test item before they could be used in the main study.
4. The test, which at this stage carried satisfied reliability value and validity index together with appropriate content validity, was transformed into test version A (native speakers) and test version B (nonnative speakers).

Data collection for the main study

1. The main study was conducted later in March 2006. The selected 192 students were assigned into two groups by using the matched pair technique, one group of 96 participants took version A test and the other took version B test. The tests lasted approximately 30 minutes and were operated in the language laboratory where the instruction usually took place. The participants were asked to take the test in groups. The participants were requested to do the two test versions in the same month.
2. During the test period, in order to control the carryover effects, the subjects took first the multiple choice format and then took the short answer format. A ten minute break was allowed after the first session.

In the second session, the short answer format came before the multiple choice format.

3. About two weeks after the participants took the test, 30 selected participants were invited for the interview. The 30 subjects were selected from each group, 15 from the version A group and the other 15 from the version B group. These 30 participants were selected according to their performance on the two test versions. They were those students who scored high, medium and low in Tests A and B. The demographic score data is presented in Chapter 4. The interview was administered individually.
4. The interview participants listened to several speakers with different accents on the stimulus CD. They were asked to respond to each voice immediately after hearing it for the attitude interview questions.
5. In the last step in the retrospective study, the interview participants were recalled. They listened to some parts of the test again to order to support their answer to the interview questions concerning their preferences towards varieties. The interview was recorded in order to be analyzed afterwards.

V. Data analysis

This study is divided into three phases according to the data analysis. The first phase is the listening comprehension test validation which involves the data analysis from the pilot study. The second phase of the study involves quantitative data in research questions 1, 2 and 3. The third phase which concentrates more on qualitative data involves research question 4. The analyses are as follows:

The First Phase: Piloting and Validating the Listening Comprehension Test

The listening comprehension test was validated to find the validity and reliability value in February 2006. The first step was to have a test trial on a small but as close as possible representative sample of the potential test population. The purpose of the first trial was to select appropriate and reliable items that have good statistical value from the test analysis. After that the test was improved before it was transformed into two versions. The steps were:

1. To validate the test content, the test items were examined by five teachers (four Thais and one native speaker) who teach the course to the participants. These teachers are considered to be the people who know the objectives of the test best since they have been teaching and writing the test for the course for many years. One Thai teacher has been teaching this Listening Comprehension course for more than 20 years. The number of five teachers is suitable because when there are two extreme different opinions towards one item, the opinion from the last teacher would be the deciding judgment for that item. A native speaker teacher's opinion is needed here in terms of correctness and appropriateness of the language used in the script. To consult these teachers, a questionnaire survey on text suitability was used together with purposes of the test and test specification. They were asked to assess on two dimensions: the suitability of conversation script used for the test and the suitability of item measurement on course objectives. The questionnaire was in the form of a checklist and it contained the objectives of the test, the item and section arrangement, and the relation of the test items to each objective, regarding 3 rates: (1) clearly measuring; (2) unclear; and (3) clearly not measuring (Appendix C). For the suitability of the conversation, the teachers evaluated the conversation on the basis of topic familiarity regarding 4 rates: (1) not familiar at all; (2) not familiar; (3) familiar; and (4) very familiar. For language difficulty, they evaluated the conversation regarding the other 4 rates: (1) very difficult; (2) difficult; (3) easy; (4) very easy, for language difficulty.
2. A feedback questionnaire was delivered to every student who tried the test to find out:
   - their perception of the language level of each conversation,
   - their familiarity with the topic of each conversation,
   - their familiarity with the response format,
   - their attitude to the response formats,
   - their opinion on time given on each section,

- their opinion on the most difficult and heaviest time pressure section.

3. After the judgment was made by the teachers, the content validity was calculated using Item-Objective Congruence Index (IOC) (Turner and Carlson, 2003.) The accepted value of each item was 0.75. The items that were below 0.75 were revised or eliminated
4. The results of feedback questionnaire from the students were calculated using SPSS to see the mean, standard deviation and percentage of each test section and conversation. The section that was reported to have an extreme result was revised.
5. To estimate the test reliability, the internal consistency was measured. The test was then tried out with 99 students who were not selected by the random sampling for the main study. These 99 subjects were chosen for the pilot study.
6. For the objective test items with the response format of multiple choice, the items were analyzed to find their item reliability and difficulty index using *Sukamolson's CTG program.* The acceptable item difficulty index is between 0.20-0.80. If the difficulty index of an item is more than 0.80, the item is very easy, and if it is less than 0.20, the item is too difficult for the students (Sukamolson, 2004).
7. For estimating reliability of the test items, Kuder-Richardson 20 (KR20) was used for the response formats of multiple choice and short answer. The acceptable value is ≥0.75 (Franenkel and Wallen, 2000).
8. For estimating reliability of the subjective test items, Cronbach's Alpha coefficient is used if each item has equal full-scores, because KR 20, KR 21 and Split-halves method are not suitable with subjective test items. Cronbach Alpha was used for the response formats of short answer. The acceptable value is ≥0.75 (Scannell and Tracy, 1975)

The Second Phase: the Main Study

1) Using a Two-Way Mixed Factorial ANOVA

In the first phrase of the study, two independent variables were involved in the design. The two variables were the listening test formats and the varieties of

English. The test format variable had 2 levels that were multiple choice and short answer format. The English variety variable had 2 levels – the native speaker stimuli and the non-native speaker stimuli. A type of the varieties of English was the randomized-group IV with two levels. A total of 192 subjects participated in the study and were randomly assigned into two groups equally. Each group listened to only one kind of listening stimuli. The test format was the repeated-measures IV which had two levels. Both groups of subjects performed on both multiple choice and short answer format. The scores obtained from both test versions served as the dependent variable. The design of the data analysis was a 2x2 mixed factorial ANOVA because it involved two factors with 2 and 2 levels respectively. A two-way ANOVA was suitable to test the significant difference of these categories. The use of ANOVA allowed the researcher to talk about the following different effects:

- The effect of test task factor (Factor A)
- The effect of the use of English varieties (Factor B)
- The effect of a combination of test tasks and English varieties (Factor A X B)

To answer research questions 1, 2 and 3 and to test hypotheses 1, 2 and 3, the scores from the two versions of listening comprehension test were calculated and compared using the two way ANOVA or *F*-ratio formulas. The data prepared were analyzed by SPSS program (release 12.0). There were six steps of calculation:

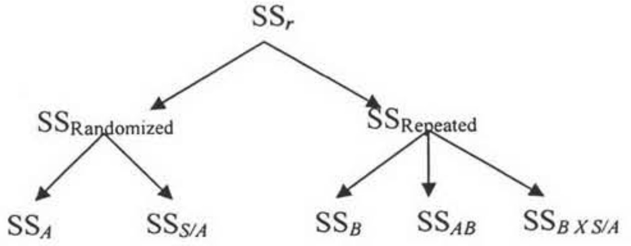
1. compute sum of squares total(SST)
2. compute sum of squares between(SSB)
3. compute sum of square within(SSW)
4. compute sum of square for Factor A(SS_a)
5. compute sum of square for Factor B(SS_b)
6. compute sum of square for interaction(SS_{ab})

A 2 X 2 (test formats and English varieties) mixed model analysis of variance were performed on the participants' listening test scores. Test version – native speaker version and non-native speaker version, was a between subjects variable and test format was a within subjects variable. An alpha level of 0.05 was used for all statistical tests.

There are things to consider about the analysis for a mixed design. Repeated measures are sticky. For instance, The A X B interaction contains both a randomized-group effect and a repeated-measures effect, but is analyzed in the repeated-measures part of the design. Any interaction term that contains one repeated-measures main effect is analyzed as part of the repeated-measures segment of the design.

Figure 3.6

Partition of Sums of Squares in a Mixed Randomized-Repeated Design



The mixed design had two IVs which were test formats and varieties of English. The test format was a randomized-group IV respresented by A. The first 96 cases were in level a_1 which refers to the native speaker version and the other 96 cases were in level a_2 which refers to the nonnative speaker version. The b_1 is the multiple choice format and b_2 is the short answer format. B represents varieties of English IV. All subjects provide a DV value at both b_1 and b_2 level, so B is a repeated-measures IV. Allocation of cases in a two-way mixed design is illustrated as:

Table 3.13

. **Allocation of Cases in a Two-Way Mixed Design**

Randomized ***Groups (A)***	***Repeated Measures (B)***	
	b_1 **multiple choice**	**b_2** **short answer**
a_1 **native version**	S_1 S_2 S_3 ... ... S_{96}	S_1 S_2 S_3 ... ... S_{96}
a_2 **non-native** **version**	S_{98} S_{99} S_{100} ... ... S_{192}	S_{98} S_{99} S_{100} ... ... S_{192}

Another thing to consider is the assumption of sphericity when an ANOVA with a repeated measures factor is conducted. Sphericity requires that the variances for each set of different scores are equal. The effect of violating sphericity is a loss of power that will increase probability of a Type II error, and the *F*-ratios produced by SPSS cannot be trusted. Mauchly's test statistic reported by SPSS can tell whether the sphericity is violated. If Mauchly's test statistic is significant ($p < 0.5$), it means that the assumption of sphericity is violated and corrections need to be done.

There are two options to deal with sphericity violation. First, there are three different corrections provided by SPSS to produce a valid *F*-ratio. They are Greenhouse and Geisser's, Huynh and Feldt's and the Lower Bound estimate. It is recommended to use Huynh and Feldt's correction if epsilon is more than 0.75 ($\varepsilon > 0.75$). However, when $\varepsilon < 0.75$ or nothing is known, Greenhouse-Geisser's correction is appropriate (Field, 2005). The second option is to use multivariate test statistic (MANOVA). Multivariate procedures will be more powerful and appropriate when there is a large violation ($\varepsilon < 0.7$) and the sample size is greater than 10+ number of levels of the repeated measures factor (Field, 2005).

2) Using an Effect Size measurement

For effective interpretation and application of research results, it is important to note one of the limitations of traditional statistical significance testing: statistical significance is highly dependent on sample size. That is the opportunities for achieving statistically significant results increase as sample size increases, and decrease as sample size decreases. Because of this dependence on sample size, "statistically significant results" cannot always be equated with "meaningful results" (Cook, 1999). For the application of research results to have the greatest impact and value for decision support needs, assessment of research outcomes should therefore not be limited to determinations of statistical significance. Consideration should also be given to the meaningfulness or practical importance of the outcome (magnitude of the outcome).

Further, ANOVA typically centers on significance, not association. However, with large samples, groups may be found to differ significantly on a dependent variable, but these differences in effect size may be small. Therefore, researchers using ANOVA are recommended to also report level of association for significant effects (NCSU, 2006).

In the first phase of the study, the 'Effect Size' measurement was used to analyze the size of the experimental effect of the test versions and the test formats (main effect), and the combination of test version variable and test format variable (interaction effect). To analyze the effect size, the researcher used two measurements of effect size value: (1) Eta squared (correlation ratio) to report association and (2) Cohen's *d* to report the magnitude of the outcome. The data were calculated and the results were interpreted as follows:

a) The simplest measure of effect size is to use Eta squared (η^2). The reason to use descriptive measure η^2 is because it easily generalizes to a variety of ANOVA designs and is often used with complicated designs (Tabachnick and Fidell, 2001). Eta squared is the percent of total variance in the dependent variable accounted for by the variance between categories formed by the independent variables. Therefore, Eta is the ratio of the between-groups sum of squares to the total sum of squares. For instance, if $\eta^2 = 0.50$, it can be interpreted that the IV and DV are associated about 50 %. The formula to calculate Eta squared used is: $\eta^2 = SS_A/SS_{Total}$ (Tabachnick and Fidell, 2001).

b) Another kind of measure of effect size expresses mean differences in standard deviation units. This is called Cohen's *d* or 'effect-size index'. Cohen (1992) shows the varieties of *d* and equations for converting η^2 to *d.* The magnitude of effect size values were analyzed using the suitable formula of Cohen's *d* for *F* test ratio. The Microsoft Excel Spreadsheet (Thalheimer and Cook, 2002) that can be used to compute the value was administered. The spreadsheet can only be used for comparing two means, in case that there are more than two means to compare at a time, the simplest formula of *d* was used: $d = 2r/\surd(1-r^2)$ (Hopkins, 2002). The Cohen's *d* value of 0.2 – 0.4 is described as 'small', 0.5-0.7 is 'medium', and 0.8 further is 'large' (Cohen, 1992).

In addition, an effect size can be directly converted into statements about (1) the overlap between the two samples in terms of a comparison of percentiles; and (2) the percent of nonoverlap of one group's scores with another group's scores. Since the effect size uses the idea of standard dev;ation to contextualize the difference between the two groups, therefore, the conversion of effect size is made possible and there is a table that shows the effect size converted to percentile standing and percent of nonoverlap available (Coe, 2000; Becker, 2000). Appendix F includes the table of the interpretation of Cohen's *d*.

The Third Phase: Retrospective semi-structured interview data analysis

The third phase of the study involves qualitative analysis. To answer research question 4, the data from the retrospective interview was analyzed qualitatively. A complete transcript of the interview was prepared from the tape recorded. Recorded interviews were transcribed into written text before being indexed. The data then was classified into taxanomy.

According to Burg (2004), qualitative data analysis can be defined as consisting of three concurrent flows of action: data reduction, data display, conclusion and verification. The data analysis process for this qualitative phase was as follows:

1. Data reduction occurred after the recorded interview was completely transcribed. It directed attention to need for focusing, simplifying and transforming raw data into a more manageable form.
2. The reduced data then was displayed as an organized, compressed assembly of information. Displays involved tables of data, tally sheets of themes or similar reduced groups of data. These displays assisted the researcher in understanding certain patterns in the data.
3. The last analysis activity was conclusion drawing and verification. After the data had been collected, reduced, and displayed, analytic conclusion began. First conclusions drawn from the patterns apparent in the data were verified by carefully checking the path of the conclusions that is to retrace the various steps that lead to the conclusions. The researcher consulted two more judges in analyzing and categorizing this script data to see if they would draw comparable conclusions. This is a kind of intercoder reliability check in order to increase the reliability of the qualitative data analysis process.

Summary

Chapter Three presents the research methodology of the study. The data of the population and sample together with sampling method are presented. The procedures employed in the development of the research instruments and the validation process of the research instruments are described. The steps taken in data collection and data analysis are also illustrated in the last part of the chapter.