CHAPTER III
RESEARCH METHODOLOGY

This chapter deals with the research procedure to explore the effects of two types of feedback – overt correction and self-correction – and language abilities of the students on the usage of English tenses. It includes the following topics: research design, population and samples, experimental materials, research instruments, data collection, and data analysis.

3.1 Research Design

The present study is a true experimental research design. The researcher used cluster random sampling technique together with random assignment when assigning samples to treatment groups. In addition, 'matching' was applied in order to increase the likelihood that the two groups of subjects were equivalent. The design is illustrated as follows:

Figure 3.1: Diagram of the research design

Self-correction	R	O_1	M_r	X_1	O_2	O_3
Overt correction	R	O_1	M_r	X_2	O_2	O_3

The symbol M_r refers to the fact that the members of each matched pair were randomly assigned to the two study groups. There were three observations which included pretest, posttest, and delayed test.

3.2 Population and Samples

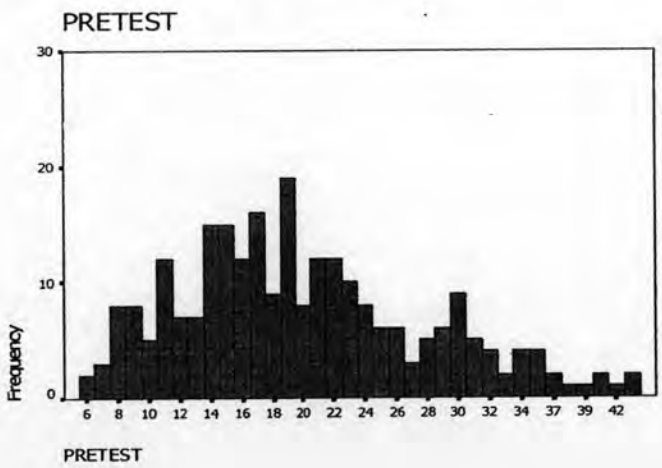
The population of this study was first-year undergraduate students in the academic year 2006 of Huachiew Chalermprakiet University, Thailand. The total number of students presented on its official website on 25th October, 2006 was 908. The population included students from 12 faculties: Business Administration, Liberal Arts, Pharmaceutical Science, Physical Therapy, Public Health, Science and Technology, Chinese Medicine, Medical Technology, Law, Social Work and Social Welfare, Nursing,

and Communication Arts. The students were taking the course *GE 1063 English for Communication II* in semester 2, 2006.

The students finished *Mattayomsuksa* 6 (equivalent to grade 12) from Thai secondary schools. They studied English at school for about 8 to 15 years before going to university. At Huachiew Chalermprakiet University, they completed the first English foundation course, *GE 1053 English for Communication I*, during the previous semester.

The optimal sample size for the population of 908 is around 270 (Krejcie & Morgan, 1970). The samples were reached by three steps. First, the *cluster random sampling* technique drew 6 groups from the 23 study sections organized by the registration office. Generally, students can enroll in any sections they want but it is common to find that students from the same faculty all study in the same section. As a result of the cluster random sampling, 279 students were selected. Second, all students were pretested in the first week. However, there were quite a lot of students who withdrew from the course in the following week because they just got the official grade report informing that they did not pass the GE 1053 which was the pre-requisite course. As a consequence, 251 students remained in the study. The frequency of pretest-score distribution of 251 students is as follows.

Figure 3.2: The pretest score distribution



It can be seen from the Figure 3.2 that the distribution is slightly positively skewed (+0.629). To divide the samples into 3 subgroups, the researcher used the 30th and 70th percentile ranks which are appropriate for skewed distribution (Tirakanant, 2003: 63). As a result, students who scored lower than 16 were put in the Low Achievers group, those who scored between 17 to 23 were labeled as Moderate Achievers, and

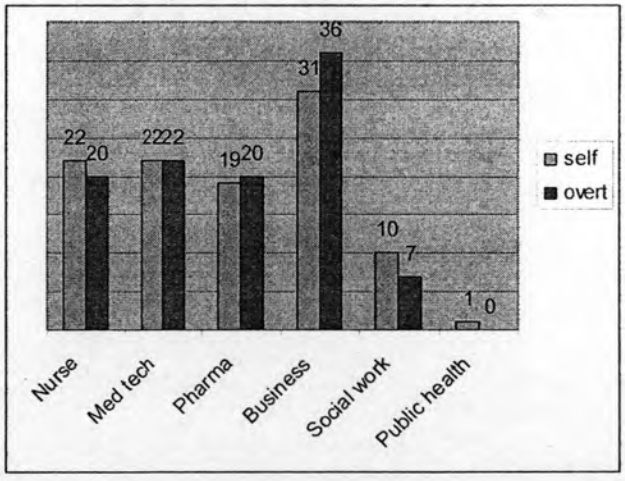
those who gained higher than 23 were labeled as High Achievers (Appendix A). Last, the samples were matched and then were *randomly assigned* to each of the two study groups — overt correction and self-correction. After excluding some students whose score did not match with any other, 210 samples remained in the study. The sample distribution was as follows:

Table 3.1: Sample distribution

		Types of Error Treatment		Total
		Overt	Self	
Language Abilities	Low	35	35	70
	Moderate	35	35	70
	High	35	35	70
Total		105	105	210

The remaining samples came from 6 different faculties as shown in Figure 3.3.

Figure 3.3: Sources of the samples



The majority in both overt and self-correction groups were from the faculty of business administration, which is in fact the biggest faculty in the university, accounting for 31.9% of the total samples. The numbers of samples from the faculty of medical technology, and the faculty of nursing were approximately the same (20.95% and 20% respectively). About 19% and 8% of the samples were from the faculty of pharmaceutical

science and social work respectively. There was only one student from the faculty of public health, accounting for 0.48%, in the present study.

In order to assure that the overt correction group and the self-correction group were comparable before the treatment, mean scores of the two groups at each ability level were compared. Descriptive figures were illustrated in Table 3.2.

Table 3.2: Mean scores, standard deviations, and ranges of scores from the pretest

	Error treatment	n	Min	Max	Range	$\bar{X}$	SD
Low Achievers	Overt	35	6	15	9	11.69	2.71
(LA)	Self	35	6	15	9	11.63	2.67
	Total	70	6	15	9	11.66	2.69
Moderate Achievers	Overt	35	16	23	7	19.29	2.27
(MA)	Self	35	16	23	7	19.29	2.32
	Total	70	16	23	7	19.29	2.29
High Achievers	Overt	35	24	44	20	30.20	4.96
(HA)	Self	35	24	44	20	30.51	5.10
	Total	70	24	44	20	30.36	5.00

It is obvious that at every ability level, the two feedback groups had similar minimum and maximum scores. This is the result of the *matching* technique used in sample assignments in order to control the variances. It is found that the HA groups had the greatest standard deviations (4.96 and 5.10), meaning that they had more variability from the central point in the distribution. This can be clearly seen from the high range of scores – both were 20. When looking at the mean scores, it can be roughly seen that the figures in each cell are not different. However, a proper statistical analysis is needed to confirm the similarities.

The researcher used an *independent-samples t-test* to compare mean scores of each ability level. The *t-test* is a fairly robust test, so users do not have to be overly concerned about normal distribution of the means (Hatch & Farhady, 1982: 114). Results from the *t-test* are presented in Table 3.3.

Table 3.3: Test of equivalent groups by *independent-samples t-test*

Ability	Error Treatment	n	$\overline{X}$	S.D.	Mean Difference	df	*t*
Low Achievers	Self	35	11.63	2.67	-.06	68	-.089
	Overt	35	11.69	2.71			
Moderate Achievers	Self	35	19.29	2.32	.00	68	.005
	Overt	35	19.29	2.27			
High Achievers	Self	35	30.51	5.10	.31	68	.261
	Overt	35	30.20	4.96			
Total	Self	105	20.48	8.56	.09	208	.073
	Overt	105	20.39	8.39			

Table 3.3 shows that the *t* values from the calculation of HA, MA, and LA groups were .261($p > .05$), .005 ($p > .05$), and -.089 ($p > .05$) respectively (see Appendix B). Besides that, the overall *t* value was not significant ($t = .073$, $p > .05$). This means that the abilities of the overt correction group (OC) and the self-correction group (SC) were not significantly different at the beginning of the study before receiving the treatment.

3.3 Experimental Materials

The experimental materials consisted of two "drill and practice" computer-assisted language learning (CALL) programs representing two different error treatments: self-correction (SC) and overt correction (OC). The development of the materials comprised two phases. Phase I dealt with the development of the experimental materials and Phase II was the implementation of the materials.

Phase I: The development of the experimental materials

The development of the experimental materials was adapted from "the model for design and development" by Alessi and Trollip (2001: 410). It comprised 3 stages:

Stage 1: Planning

1.1 Resource

1.2 Learner characteristics

1.3 Related theories

1.4 Feedback options

1.5 Selection of contents

Stage 2: Design

2.1 Content ideas

2.2 Scripts and the content validity

2.3 Flowcharts and storyboard

Stage 3: Development

3.1 Creation of the 'prototype'

3.2 Alpha testing and revision

3.3 Beta testing and revision

Stage 1: Planning

At this stage, the researcher explored the availability of necessary resources at the university, identified learner characteristics, and studied concepts related to the design of the CALL software. The concepts include advantages and limitations of CALL, theoretical frameworks that are related to the design, feedback options, and the selection of contents. This is essential information that would be used as the foundation for designing the CALL.

1.1 Resources. It was found that there were eight computer laboratories which were almost always used by the faculty of Science and Technology and the faculty of Business Administration for their regular classes. However, students could also use computers at the computer center located in the library building. All computers were equipped with enough memory and support systems to run the program, except the sound devices. This was the limitation of the resources, resulting in absence of listening tasks in the program.

1.2 Learner Characteristics. The learners were first-year undergraduate students of a private university whose English language ability varied from low to upper intermediate. They were between 17 to 20 years of age. The learners came from any of the 12 faculties in the university, so they might have slightly different motivations and background knowledge.

1.3 Related Theories. The present study was aimed at comparing overt correction and self-correction by developing two CALL programs to represent the two error treatments. Overt correction (OC) program represented the traditional 'teacher correction' in which the teacher would indicate that the answer is wrong, explain, and

give the correct answer. The self-correction (SC) program, on the other hand, would not give the correct answer to the students. Students were encouraged to find the correct answer by themselves. The researcher has adopted and has applied the concepts of a number of theoretical frameworks–constructivism, noticing hypothesis, and autonomous learning (for details, please read 'the theories concerning SC and OC' in chapter II). The researcher has also adopted the concepts of 'scaffolding' when planning for CALL tasks. Scaffolding involves the setting up of "temporary supports, provided by capable people, that permit learners to participate in the complex process before they are able to do so unassisted" (Peregoy & Boyle, 1997; cited in Ediger, 2001: 166). The SC program gave specific feedback and support in steps (Table 3.4), until the students could get to the correct answer.

Table 3.4: Steps in providing feedback in the self-correction program

If the student chooses an incorrect answer for ...	The program will...			
	Indicate that the answer is wrong with "X" sign	Provide specific feedback why the choice is wrong	Highlight the keyword(s)	Provide an explanation for the situation in the Thai language
The 1st time (3 choices left)	✓	✓	✓	
The 2nd time (2 choices left)	✓	✓	✓	✓
The 3rd time (1 choice left)	✓	✓	✓	✓

* The specific feedback given at the first, second, and third trials are different depending on the error of the selected current choice.

If the student chooses the wrong answer the first time, the program will first indicate that the answer is incorrect. Specific feedback explaining why it is incorrect will appear. These two options are set as a program standard for every response. The first step of 'scaffolding' is to highlight the keyword(s) or the clue(s) given in that item. Examples of keywords are a specific adverb of time and the tense used in the leading questions. Then the student will have to think and make a decision again. If he/ she selects the wrong choice the second time, apart from the standard options and the highlighted keywords, explanation of the situation in Thai will be supplied. Again, the student will use all of the support to make his/ her judgment for the third time. At this stage, there are only two choices left. With such help, he/ she should be able to detect the correct answer. However, if he still cannot get it right, the program will give the same treatment as given for the 2nd trial, except that the specific feedback will be changed to correspond to the choice selected.

1.4 Feedback Options. Kulhavy (1977; cited in Boonplong, 1998: 37) reviews studies about error feedback and suggests that negative feedback seems to be more effective than positive feedback. It is also suggested that positive feedback should be provided after the students have started to work for a while or they finish half of the lesson. It is not necessary to give positive feedback every time they answer correctly.

Cohen (1985; cited in Boonplong, Ibid.) gives suggestions for CALL developer that:

1. The program should not give "praising" feedback;
2. The program should provide immediate feedback;
3. Feedback should be always provided no matter wheter the answer is correct or not; and
4. Feedback should contain explanations telling the student why his answer is correct or why it is incorrect.

Alessi and Trollip (2001: 115) state that feedback should provide the learner with information to improve future performance. It should avoid negative statements, sarcasm, and should never demean the learner. Also, it should be noted that providing highly *interesting* feedback may increase the rate of errors. When feedback following errors is much more interesting than that following correct responses, the learner may be stimulated to make errors intentionally to see the interesting effects.

Research regarding timing of feedback shows that immediate feedback is not always more beneficial than delayed feedback, but is almost always better than no feedback (Alessi & Trollip, 2001: 115). The proper timing depends on the nature of what is being learned and how it is being learned. In general, an advantage for delayed feedback has been demonstrated for propositional knowledge (verbal information, knowledge, principles). In contrast, immediate feedback is more likely to enhance learning procedural knowledge (Anderson, 1982; cited in Alessi & Trollip, Ibid.).

Application to the present study: The researcher planned to use both positive and negative feedback. Positive feedback would be given when the students answer correctly, while negative feedback would be used with incorrect answers. As suggested by Cohen, the feedback would be always provided for both correct and incorrect answers. Moreover, explanations would be given to students as to why the answer is correct or incorrect.

Regarding the timing of the feedback, the researcher decided to make use of immediate feedback because the learning of verb tenses is regarded as procedural

knowledge rather than propositional (or declarative) knowledge. According to Nunan (1999: 305), declarative knowledge is the knowledge that can be stated, such as grammatical rules. In contrast, procedural knowledge has to do with the ability to use the knowledge to get things done, for example, being able to use grammatical rules and principles to communicate meaning. The students in the present study were expected to be able to use the tenses appropriately rather than to state the grammatical rules. Therefore, immediate feedback was applied to the CALL.

1.5 The Selection of Contents. Regarding the contents of the materials, Wingfield (1975; cited in Hendrickson, 1979: 16) has suggested that self-correction would probably be effective with **grammatical errors** but would be relatively ineffective with lexical errors. This claim is consistent with the findings from Zhao's study (1997; cited in Sukamolson, 2000: 31). Zhao investigated attitudes of directors of Intensive ESL programs towards the use of CAI in American universities and found that the participants agreed that CAI was an appropriate tool in language teaching. It was more suitable for beginners than advanced students. Moreover, the directors thought that CAI was appropriate for teaching grammar rather than teaching reading or writing. Therefore, the CALL contents were developed around common grammatical problems of Thai students. Smyth (1987) proposes that plurals of nouns and verb tenses are normally unmarked for Thai speakers of English. This claim is supported by Leong's study of (1980) with Thai students at Chiengmai University. Leong found that plurals, subject-verb agreement, and tense are the most frequent grammatical errors made by the students accounting for 26.1%, 22.1%, and 13.9% of total errors, respectively. Similarly, from her study entitled 'A study of error analysis in English compositions: case study of English major students of Rajabhat Institute Nakorn Pathom', Khaourai (2002) reported that grammatical errors were the most frequent error type found in the study. Among nine categories of grammatical errors, 'tenses' was found to be one of the most frequent errors in both guided writing (rank 4, 12.31%) and in free writing (rank 1, 30.38%). Based on the findings from the mentioned studies, the researcher has chosen "**tenses**", which is one of the most problematic grammatical categories, as the contents of CALL. Moreover, the topic was relevant to the contents of the foundation course that the subjects were taking.

The term *tense* refers to the full name of the verb form, which always consists of two elements: a time and an aspect (Master, 1996: 46). There are three times that are expressed by English grammar: past, present, and future. Aspect comprises simple, continuous, and perfect. Each verb tense in English consists of one of the three times and one or more of the three aspects. The twelve tenses in English are listed in Table 3.5.

Table 3.5: The twelve English tenses

Aspect	Time		
	Past	**Present**	**Future**
Simple	**Simple Past**	**Simple Present**	**Simple Future**
Continuous	**Past Continuous**	**Present Continuous**	Future Continuous
Perfect	**Past Perfect**	**Present Perfect**	Future Perfect
Perfect Continuous	Past Perfect Continuous	Present Perfect Continuous	Future Perfect Continuous

Master suggests that seven of the twelve tenses are used much more frequently than the remaining five. They are: 1) Simple Past, 2) Simple Present, 3) Simple Future, 4) Past Continuous, 5) Present Continuous, 6) Past Perfect, and 7) Present Perfect. By examining textbooks used in English foundation courses,* it is confirmed that these tenses are introduced to learners in Foundation courses. The contents of the experimental materials were then to focus on these seven tenses.

Stage 2: Design

After obtaining essential information from the first stage, at this stage, the researcher developed initial content ideas, prepared scripts, and created flowcharts and storyboards.

2.1 Content ideas. The experimental materials or CALL, comprised 9 modules– 6 lessons and 3 tests. In each lesson, two related tenses were compared and contrasted (see Table 3.5). The sequence of the presentation is based on an assumption of the different difficulty levels of each tense obtained from the review of textbooks written by scholars in the field. The two tenses that have been compared and contrasted are found related in some ways. For example, the first module present the simple present tense as contrasted to present continuous tense because these two are considered the easiest and both talk about 'present' time. Next, when students understand these two fundamental tenses, the simple past tense is inserted. The reason behind comparing the simple present tense to the simple past tense is based on the evidence found in Khaourai' s (2002) study. In her study, Khaourai reported that among the errors regarding tenses, Thai undergraduate students used present simple tense instead of the past simple tense most frequently in both free writing and guided writing. These two tenses are then put

* Examples of the textbooks included: Transitions (Lee, 1999), True Colors (Maurer & Schoenberg, 1998), and Interactions (Baldwin et al, 2003).

together. The misuse of past simple and present perfect was also reported in the same study. They are then put as the contents for the following module. The rest of the contents are sequenced in accordance with the scholars mentioned earlier. The sequence was validated and agreed upon by experts who are experienced English language teachers (for expert profile, see appendix D).

Table 3.6: Sequence of the CALL contents

Module	Contents
1	Simple Present Tense vs. Present Continuous Tense
2	Simple Present Tense vs. Simple Past Tense
3	Test 1 (3 tenses in lessons 1 & 2)
4	Present Perfect Simple Tense vs. Simple Past Tense
5	Future Tenses (Future Simple, Present Continuous, and be going to)
6	Test 2 (5 tenses above)
7	Past Continuous Tense vs. Simple Past Tense
8	Past Perfect Simple Tense vs. Simple Past Tense
9	Test 3 (all 7 tenses)

The three tests were similar to the six lessons in terms of the error treatment they provided. However, the coverage of contents of the tests was broader. Test 1 combined contents from lesson 1 and lesson 2, meaning that it covered three tenses. Test 2 covered five tenses and test 3 mixed up all seven tenses together. Another function that was added to the tests was "score" presentation in order to help students evaluate their progress in each particular practice area. On the last screen of the tests, there was a summary of score presented both as raw data and as a percentage.

The items were in multiple-choice and matching formats. The content difficulty started from simple and concrete contents at the beginning to longer and more complicated contents at the end of each module. There were 20 items in each module. Students could take around 10-30 minutes to complete each module, depending on individual pace of learning.

At the bottom of each page on the monitor, there were 4 buttons that linked to 4 sets of content pages – forms, usage, keywords, and glossary. Students could open the links and see the contents at any time during the practice.

2.2 Scripts and the content validity. At this stage, the researcher prepared scripts of the contents (Appendix C). Regarding the quality of the materials, the contents were validated by 3 experts who are experienced English language teachers (for expert

profiles, see Appendix D). The evaluation form consisted of two parts. The first part regarded the contents, while the second part was concerned with the feedback (Appendix E). There were 6 and 5 items in each part respectively. Therefore, the whole evaluation form comprised 11 items that were presented in the form of 5-point numeral Likert-type scales:

5 = very good
4 = good
3 = acceptable
2 = poor
1 = needs work

Experts were asked to rate from 1 to 5 according to the extent to which they agreed with each statement. The evaluation criteria of the validation form were as follows:

0.00 - 1.50 means that the contents of the CALL was of 'very low' quality.
1.51 - 2.50 means that the contents of the CALL was of 'low' quality.
2.51 - 3.50 means that the contents of the CALL was of 'acceptable' quality.
3.51 - 4.50 means that the contents of the CALL was of 'good' quality.
4.51 - 5.00 means that the contents of the CALL was of 'very good' quality.

The average score of each item is shown in Table 3.7.

Table 3.7: The validation of the CALL contents

No.	Traits	Expert A	Expert B	Expert C	$\overline{X}$	Grand mean
	Contents					
1	Content accuracy	4	4	5	4.33	
2	The difficulty suits levels of the learners	5	5	5	5	
3	Content Coverage	4	4	5	4.33	4.67
4	Sequence of all the lessons	4	5	5	4.67	
5	Sequence of the items in each lesson	5	5	5	5	
6	Enough numbers of the items per lesson	5	4	5	4.67	
	Feedback					
7	Correctness of the feedback	5	4	5	4.67	
8	The feedback is comprehensible	5	4	5	4.67	
9	Feedback order	5	5	5	5	4.80
10	Specific to particular errors	4	5	5	4.67	
11	Usefulness to learners	5	5	5	5	
Grand mean for all 11 items = 4.73						

The average scores ranged from 4.33 to 5.0. The grand means of the content and the feedback sections were 4.67 and 4.80 respectively while the grand mean of all items was equal 4.73. This can be interpreted that the experts rated the overall contents and the feedback as of very good quality. Apart from being asked to rate their opinions towards the contents and the feedback, the experts were also asked to give additional comments and/ or suggestions concerning ways to improve the program. Comments and/ or suggestions from the experts were as follows:

Expert A: "The program provides personalized feedback and scaffolding that will help learners as they go along. This will be really helpful when teachers have a big class. When the program is integrated in the classroom and used in combination with other kinds of meaningful activities, learning should be facilitated. With regard to the test items, please make sure that the feedback for each error is clear and specific enough. As for the content, it's nice that two tenses are presented to show students how each tense is distinct. One question, though, is there any reason why you didn't include the perfect continuous aspect?"

Expert B: "For weak students whose English proficiency is low to very low, they may not be able to read and understand the feedbacks, not mentioning about the explanation parts like keywords/ clues, etc. Therefore, the Thai explanations should be of some help for their understanding. Alternatively, if they are explained about some key phrases or vocabulary items such as 'does not exist', 'habitual activity', or 'the base form', this may help."

Expert C: "The content is suitable for beginners. Each grammatical form and usage (affirmative and negative statements, yes/no and wh-questions) is presented in clear and easy-to-understand charts. The combination of the grammar charts and grammar notes provides a complete reference guide for the students. The practice section provides enough varied controlled exercises with clear and concise feedback for individual study.

The researcher took all of the suggestions into consideration when developing the program. As suggested by expert B, explanations in Thai and a glossary were added accordingly."

2.3 Flowcharts and storyboards. A flowchart is a chart or diagram of how the program progresses or flows. It is a useful tool for designers to analyze program components and their sequence for their own understanding, and for communicating that

information to programmers and other designers (Alessi & Trollip, 2001: 503). The flowcharts of the CALL are illustrated in Figure 3.4 and Figure 3.5.

Figure 3.4: The CALL flowchart

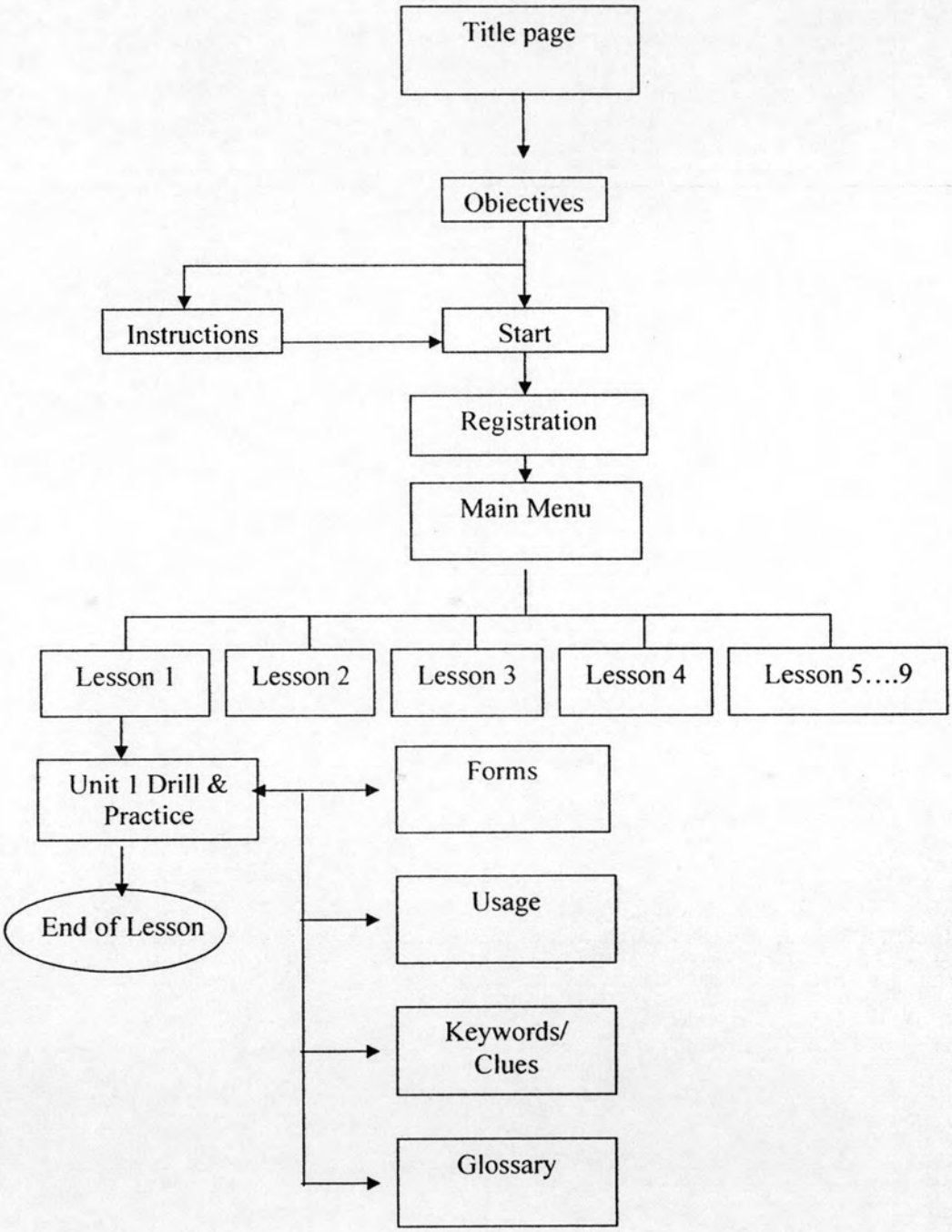


Figure 3.5: Structures and flows of each CALL item

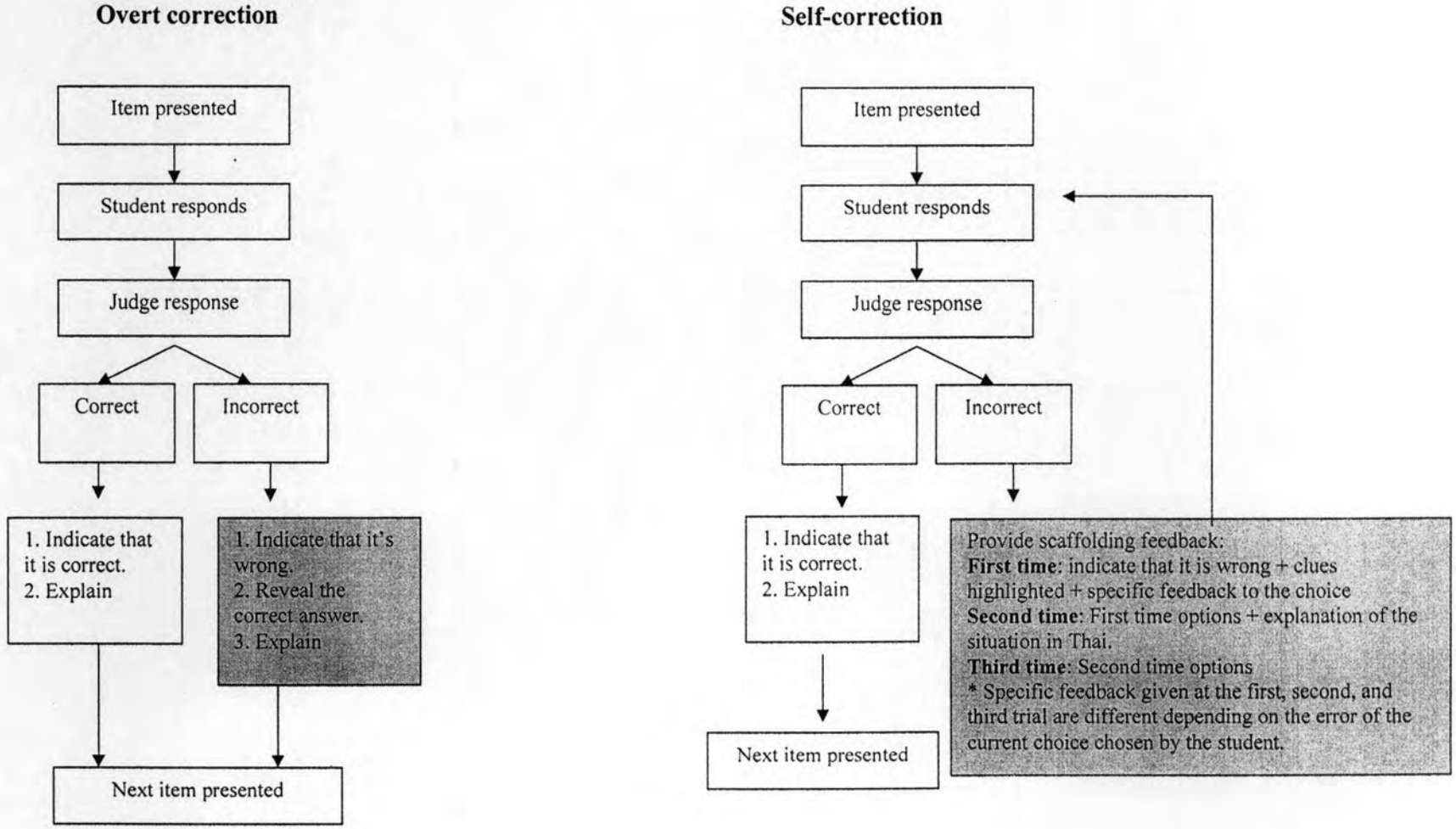


Figure 3.4 depicts the program from beginning to end. The program started from the title page, then, moved to the objective page. There were two options for them to choose on this page – to go to the instruction page or to start the drills. Before they could start, they were required to register on the registration page. After that, the students could get to the main menu that provided links to the nine modules. They were asked to go in sequence. Every lesson had 4 buttons that linked to 4 supporting contents– forms, usage, keywords/clues, and glossary.

Figure 3.5 provides details of the flow of each item in the OC and SC programs. As stated earlier, the only one difference between the two programs is how they treat incorrect answers. It can be seen from Figure 3.4 that the OC program allows students to make a decision only once, then, whether the answer is right or wrong, they can move on to the next item. In contrast, the SC program allows students to go to the next item only if they can detect the correct answer by themselves.

Flowcharts are a bird's-eye view showing the *structure* and *sequence* of the program, whereas storyboards show the details of what learners *see* (Alessi & Trollip, 2001: 503). Storyboards provide a visual representation of the design, as well as most of the details. The original storyboards were written by hand. Examples of the storyboards are shown in Appendix F.

Stage 3: Development

At this stage, the concepts from stage one and the design from stage two were implemented. First, the prototype of the first 3 modules was created. Then, alpha testing was done, followed by a revision. Last, beta testing was conducted. The researcher improved the programs for the last time before using them in the main study.

3.1 Creation of the prototype. The prototypes of the first 3 modules– Lesson 1, Lesson 2, and Test 1 – were created. This accounted for 33% of the total content. The materials were developed in the form of stand-alone programs by using *Macromedia Flash* package. Therefore, there was no need to download the 'program player' to all computers.

3.2 Alpha testing and the revision. After the prototype was developed, it was primarily evaluated by the research co-adviser, who is an expert in the field of instructional design. After the initial revision, the programs were evaluated by external experts. With regards to program testing, Alessi and Trollip (2001: 548) suggest that most projects should have at least two major tests-alpha and beta tests. The 'alpha' test is

the major test of the program by the design and development team in order to identify and then eliminate as many problems as possible. The 'beta' test is done by the client.

In alpha testing (generally for commercial software), the production staff, the instructional designers, and content experts are asked to go through the program to evaluate the content, the flow through the material, the robustness of the programming, and so on (Alessi & Trollip, Ibid.). The present study conducted alpha testing by having 3 experts who are experienced instructional designers evaluate the program (see Appendix D for expert profiles). The evaluation form (Appendix G) consisted of 3 parts. Part I contained 8 items concerned with the design in general. Part II consisted of 4 items regarding pedagogical aspects. Part III comprised 2 items that asked about the robustness and the data recording system. The items were presented in the form of 5-point numeral Likert-type scales:

5 = very good
4 = good
3 = acceptable
2 = poor
1 = needs work

Experts were asked to rate from 1 to 5 according to the extent to which they agreed with each statement. The evaluation criteria of the validation form were as follows:

0.00 - 1.50	means that the design of the CALL was of 'very low' quality.
1.51 - 2.50	means that the design of the CALL was of 'low' quality.
2.51 - 3.50	means that the design of the CALL was of 'acceptable' quality.
3.51 - 4.50	means that the design of the CALL was of 'good' quality.
4.51 - 5.00	means that the design of the CALL was of 'very good' quality.

The average score of each item is shown in Table 3.8.

Table 3.8: The validation of the CALL design

No.	Traits	Expert D	Expert E	Expert F	$\bar{X}$	Grand mean
	Design, Interface & Navigation					
1	Displays	4	5	4	4.33	
2	Screen Design	5	4	5	4.67	
3	Instruction Design	4	4	4	4.00	
4	Consistency	4	5	5	4.67	4.29
5	User control	4	5	5	4.67	
6	Directions	4	1	5	3.33	
7	Text Quality	5	4	5	4.67	
8	Navigation Aids	4	4	4	4.00	
	Pedagogy					
9	Motivation	3	4	4	3.67	
10	Interactivity	4	5	5	4.67	4.34
11	Format of Feedback	4	5	5	4.67	
12	Quality of Feedback	4	5	4	4.33	
	Others					
13	Records and Data	-	1	5	3.00	3.67
14	Robustness	5	4	4	4.33	
Grand mean of all 14 items = 4.22						

The grand mean of all 14 items was equal 4.22, meaning that the experts thought that the materials as a whole were of 'good' quality. Seventy-nine percent (11 out of 14 items) of the items had average scores between 4.33 to 4.67. There were two items (item 6 and item 13) that were rated '1- needs work' by one of the experts. Item 6 which asked about the quality of the 'directions' of the program was rated 1 by expert D because the 'directions page' was not fully developed at that time. There was only the button that linked to the page; the information describing directions had not been filled yet. However, such information was later put on the page.

As for item 13 that asked about the 'records and data' system of the materials, expert D did not evaluate it, while expert E rated it 1. This was because the data were recorded on a different database that the experts could not reach. Therefore, expert D stated that "it was unclear how the program records and reports the score in each lesson to the learner". The researcher had a chance to explain this quandary to expert D how to get such information when picking up the evaluation form.

There were other comments/ suggestions by the experts as follows:

Expert D: "1) The program requires the users to finish all the questions so that they can quit. However, in reality, users should be able to stop the program at any time they want. So, there should be an 'exit' button as an option for users. 2) After the drill, it would be nice if the students can see their scores. 3) To increase the motivation of the learners, you should add more comics or pictures to the frames."

Expert E: "After finish each lesson, the screen should show the menu of 9 lessons before quit out the program."

Expert F: "1) There should be more explanations on the usage of parts of speech. When the student answers, whether it is correct or not, there should be comprehensible explanation together with a clear example. 2) The researcher should have a course syllabus, a lesson plan, and a session plan. There should also be a satisfaction and attitude check list. 3) The program should allow learners to get back to see the previous items. 4) The review of related learning theories would be beneficial and would affect the design of the program. 5) The record system should be able to record individual information, e.g. performance, time-in, time-out, time used when reading the content pages. The data would be useful to prepare a learning guidance for individual student."

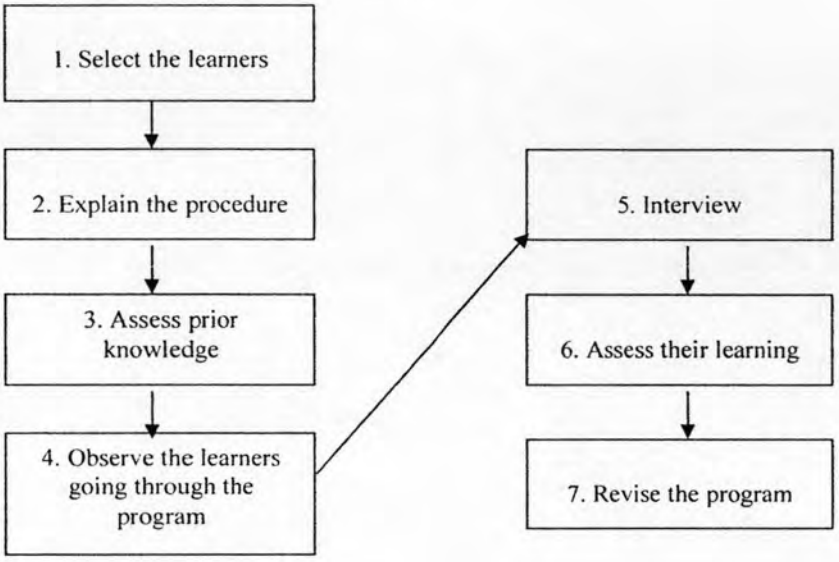
Revision: In general, the programs were considerably good. More pictures and cartoons were added as suggested. With regards to the suggestions on the flow, the researcher agreed that learners should be able to get back to the previous items and should be able to quit the program at any desired time. However, the programs would be initially used as research tools to collect data in the present study. Thus, the researcher had to control the flow in order to minimize extraneous variables. Students were purposely not allowed to quit until they finished each module. Also, they were not allowed to go back to see or to do the previous items again in order to control the amount of exposure to the drill.

Regarding the suggestions from expert F, the researcher did the review of related learning theories as presented in Chapter II under topic 3.3 CALL and the learning principles. Also, explanations were provided to students in both OC and SC programs. At the time that the expert validated it, the SC program already had the explanations given in steps; it was not of the concern. The researcher added the 'explanation' function to the OC program. So, a full explanation of how to get to the correct answer is supplied as a response no matter wheter the answer is right or wrong.

3.3 Beta testing and revision. Beta testing is another type of program evaluation suggested by Alessi and Trollip (2001: 550). It is a formal test of a final product by the

client. Alessi and Trollip (Ibid.) recommend a seven-step process to conduct a beta test as presented in Figure 3.6.

Figure 3.6: Steps in beta testing

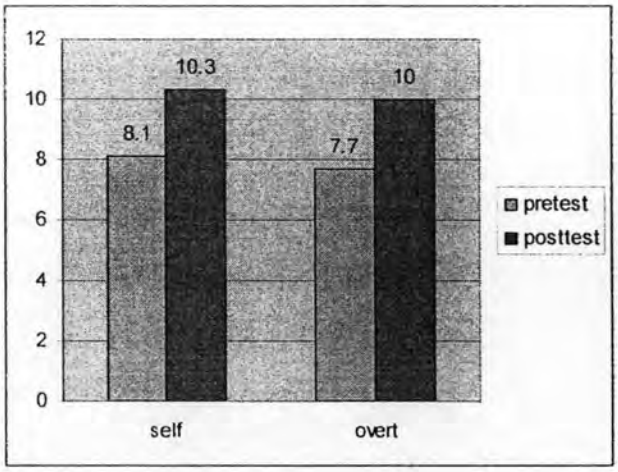


The first step was to select learners who had the characteristics of the program's target users. The target population of the present study is first-year undergraduates of Huachiew Chalermprakiet University. The researcher then picked 20 first-year students who were not in the sample groups to do the pilot test. Ten students used the self-correction program and another ten did the overt correction. They were told the purpose and procedure and then were tested for prior knowledge. Contents of the pretest covered only 3 tenses in the prototype. There were 15 multiple-choice items on the test. Then, they were allowed to practice with the programs. Observations by the researcher were made as they went through the programs. After they finished with the programs, they were interviewed and were tested again to assess their learning. Results from the posttest were presented in Table 3.9 and in Figure 3.7.

Table 3.9: Results from the beta testing

		N	$\overline{X}$	SD	Gaining Average
Pretest	Overt	10	7.70	1.77	
	Self	10	8.10	1.66	
Posttest	Overt	10	10.00	1.49	29.87 %
	Self	10	10.30	1.57	27.16 %

Figure 3.7: Pretest and posttest means from the beta testing



It could be seen from descriptive statistics that the students in both overt and self-correction groups gained higher scores from the posttest than from the pretest. Mean score of the overt correction group increased 2.3 points or 29.87 % from the pretest score. Similarly, students in the self-correction group gained an average of 2.2 points, accounting for 27.16 % of the pretest score. Figure 3.6 illustrates the mean differences. Although these figures could not be used to assure program effectiveness because of the small numbers involved, it could be said that the programs were 'likely' to be successful in promoting gains on the topics.

Apart from the concrete figures, the researcher made observations and interviewed the students. From the observations, the students did not show signs of having any problems when using the programs. From the structured interview, all students reported favorable attitudes about the programs. They thought it was fun and easy to use; the explanations were easy to understand; the program was useful and good in that they could learn by themselves. Three of the students said that they felt like they were "playing games" while practicing. However, there were some comments and requests from the students. One student thought there should be more items in each module while another wanted to have explanations in Thai. He said he had difficulty reading the feedback that was in English. Last, students wanted to see more graphics and pictures.

Revision: Once the researcher obtained data from beta testing, further revision was made. More related pictures and explanations in Thai were added. However, the researcher did not add any more items to the modules because it would contradict the

theory saying that it is more tiring to read from the screen. Moreover, if the modules were too long, students would be less motivated to finish them. Results from the pretest and the posttest confirmed that the programs, to some extent, helped improve scores. Therefore, after the final revision, the programs were good enough to be used as research materials.

Phase II: The implementation of the materials

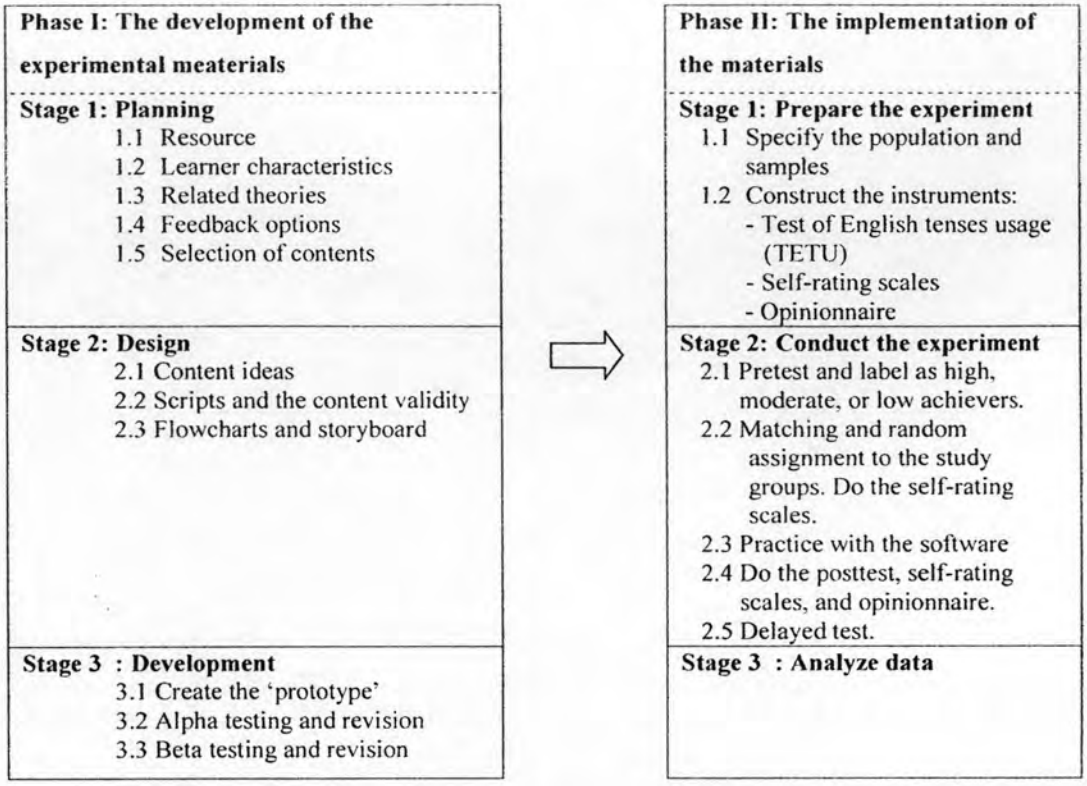
After the materials were developed, they were implemented in the study. The implementation consisted of 3 stages. Stage one dealt with the preparation for the experiment. At this stage, the researcher specified the population and the samples and constructed research instruments (for details, see topic 3.2 *population and samples*, and 3.4 *research instruments*). Since the present study had to be involved with a large number of students and needed their co-operation to answer many instruments, students who completed all research activities were awarded the extra 10% of the course evaluation.

At the second stage, the experiment was conducted. The samples were pretested, labeled, and then were randomly assigned to the two study groups – self-correction and overt correction. The researcher was allowed to use the last 0.5 to 1 hour of the regular class time to collect data, for example to conduct tests, self-rating scales, and opinionnaire. However, the practice of the CALL did not appear in the regular class time. The students could come to the university computer lab which was reserved for this project and practiced there or they could do it on-line at home if the lab time did not match their schedule. The researcher monitored their practice by checking results recorded on the server and met them at the regular class time every week during the practice period to discuss any difficulties that the students might have and to control them to strictly follow the research schedule. The conduct of CALL lessons took three weeks. After they completed all nine modules, they were posttested. The delayed test was administered 6 weeks after the posttest (for details, see topic 3.5 *data collection*).

During the third stage, the researcher analyzed the data obtained from the research instruments in order to answer the research questions. Several statistical analyses were applied (for details, see topic 3.6 *data analysis*).

The development and implementation of the experimental materials are summarized and illustrated as in Figure 3.8.

Figure 3.8: The development and implementation of the experimental materials



3.4 Research Instruments

Research instruments refer to the data collecting tools. There were 3 research instruments used in the present study: Test of English Tenses Usage (TETU), self-rating scales, and an opinionnaire.

1. The Test of English Tenses Usage (TETU). The test was developed by the researcher to serve in pretesting and posttesting. Its contents cover the seven English tenses that the subjects have practiced through the CALL programs. The test comprises 2 parts and starts from simple to more complicated tasks. Details of the tasks are presented in Table 3.10.

Table 3.10: Descriptions of the TETU

Parts	Level	Format	Total items/ scores
1. Sentence completion	Sentential level	* Discrete point test * Multiple-choice/ 5 alternatives	30 items / 30 points
2. Grammar in context	Semi-discourse level	* Context given * Longer conversations * Mixed tenses * Multiple-choice/ 5 alternatives	20 items / 20 points

The first part was a *sentence completion task* which is suggested by VanPatten and Sanz's study (1995; cited in Sanz & Morgan-Short, 2004). This part consisted of 30 discrete-point items with 5 alternatives. The item had a blank for which the students were asked to choose the appropriate verb form to fill in (see Appendix H).

The second part was called *grammar in context* because the contexts were given together with the items. Normally, items in this part came in conversations with short texts explaining the contexts. The contents were adapted from commercial textbooks written by native speakers of English. The conversations covered the use of mixed tenses. There were 20 multiple-choice items in this part.

The structure of the CALL and that of the TETU were similar as they both used multiple-choice format. Besides that, the sequence of tasks was the same. Both started with short and concrete items and ended with more complicated and mixed-tense tasks like completing conversations or letters. However, it should be noted that all of the CALL contents were not the same as those in the TETU. In other words, the TETU tested the same tenses but not the same examples. A table summarizing the items that tested each tense is provided in Appendix H.

The TETU was used three times, in the pretest, posttest and delayed test. The researcher reshuffled the items and made slightly changes on the names of persons or places on the test. Nevertheless, the contents of the three tests were the same.

Scoring: For both parts, the scoring system was very objective. Students would receive one point if they chose the correct answer, or zero if they picked up the wrong choices. Total possible score for these parts was 50.

Mode: The TETU was conducted through a traditional paper and pencil mode in order to avoid the subjects using any help from the computer.

Validity and Reliability of the TETU: The content validity of the test items was evaluated by 3 experts in the field of language teaching and testing (see Appendix D for expert profiles). The experts were asked to rate each item as to whether it was congruent with the objective stated. Then, the Item-Objective Congruence (IOC) Index was calculated by assigning scores to the answers as follows:

Congruent = 1
Questionable = 0
Incongruent = -1

The IOC index ranges from -1 to 1. Items that have an index lower than 0.5 should be improved (Tirakanant, 2003: 140) and the overall content validity index should be ≥ 0.75 (Sukamolson: 1995). Overall, the content validity of the test was 0.81. The

value of IOC for each test item was illustrated in Appendix I. Results indicated that 88 % of the items were rated higher than .05 of the IOC index, meaning that they were acceptably congruent with the objectives. Only 6 items needed improvement. The comments mostly centered on the alternative choices. The test was improved accordingly.

After the improvement, the test was piloted with 129 students in February and March, 2006. The students were in the first year and had similar characteristics to the subjects in this study. After the administration of the test, all test items were analyzed by a computer program named Siree-ER (Vatanavigkit, 2001). *Item analysis* was conducted to find the reliability, difficulty indexes, and discrimination indexes of the test (Appendix J). The reliability of the overall test calculated by Kuder-Richardson- 20 formula (KR-20) was 0.90, which can be interpreted as meaning that the test had 'high' reliability (Vatanavigkit, Ibid.). The average difficulty index and the average discrimination index of the whole test were 0.50 and 0.40 respectively. The criteria for these indices were set as follows (Sukamolson, 1995: 31):

For the difficulty index (p):

$p < 0$ .20	means the item was difficult.
p = 0.20-0.80	means the item was good in terms of its difficulty.
p = 0.81-0.94	means the item was easy.
$p \geq 0.95$	means the item was very easy.

For the discrimination index (r):

r = 0	means the item had no discrimination ability.
$r \leq 0.19$	means the item had a low discrimination ability.
r = 0.20-0.29	means the item had a fair discrimination ability.
r = 0.30-0.39	means the item had a high discrimination ability.
$r \geq 0.40$	means the item had a very high discrimination ability.

According to the criteria, the test items of which difficulty indices ranged between 0.20 and 0.80, and discrimination indices were equal or higher than 0.20 were chosen for the main study. There were 41 items that were satisfactory and 9 items needed improvement (see details in Appendix J). The researcher utilized results from item analysis and the test alternatives were improved accordingly.

2. Self-rating scales. The self-rating scales were developed for eliciting the students' perception of their own learning. Each individual student was asked to rate their knowledge about the seven tenses. The instructions were in Thai. The scales were 5-point scales representing the quality of knowledge as follows:

4 = excellent
3 = good
2 = average
1 = poor
0 = don't have the knowledge at all

There were 7 items that posed questions about 7 tenses (Appendix K). Apart from the names of the tenses, the researcher provided an example which was a full sentence using the particular tense. This helped prevent the mismatching of names and forms. The students were asked to do this twice, once before the treatment, and again after the treatment so that the researcher could compare the responses to see if the students thought they gained knowledge from the programs. The information here could be used to back up results from the opinionnaire items that asked about attitudes towards the programs. The quality of the self-rating scales was assessed by Cronbach's alpha formula via the SPSS package. The pre-treatment and post-treatment scales showed high reliability both at .90.

3. The Opinionnaire. The opinionnaire consisted of two parts. The first part asked for personal information of the students. The second part was designed to explore students' opinion towards the program that they used (Appendix L). It comprised 13 items that were presented in the form of 5-point numeral Likert scales. Students were asked to rate from 1-5 according to the extent to which they agree with each statement:

5 = Strongly agree
4 = Agree
3 = Average
2 = Disagree
1 = Strongly disagree

The evaluation criteria of the opinionnaire were as follows:

0.00-1.50	means positive opinion towards the program was 'very low'.
1.51-2.50	means positive opinion towards the program was 'low'.
2.51-3.50	means positive opinion towards the program was 'moderate'.
3.51-4.50	means positive opinion towards the program was 'high'.
4.51-5.0	means positive opinion towards the program was 'very high'.

The items could be grouped into 4 categories. The first category (items 1-6) asked about the contents, design, and the interface. The second category (item 7 and item 9) asked about the students' attitudes towards the program that they were using. The third category comprised item 8 and item 13 that asked about the students' attention when practicing. The last category asked for the students' opinions towards the feedback they received. This category consisted of 3 items, items 10-12. All items were printed in Thai. The opinionnaire was distributed to the students at the end of the study. By using Cronbach's alpha formula via the SPSS package, the opinionnaire showed high reliability at .85

3.5 Data Collection

The main study was conducted in semester 2, academic year 2006. The process of data collection was as follows:

Week 1: It was explained to all subjects that they had been chosen to participate in a research project aimed at evaluating an English software program. Then they were pre-tested. The test was one hour long.

Week 2: By using scores from the pretest, students were labeled as High, Moderate, or Low Achievers. After matching the scores, the subjects were randomly assigned to 2 feedback groups– overt correction or self-correction. In the classroom, the researcher gave an orientation for the subjects, and then gave them the CD and the checklist. The checklist was designed to help them follow their practice schedule. It stated the time frame and the modules that the subjects should practice during the time. The subjects were also asked to answer the self-rating scales in this week.

Weeks 3-5: The subjects practiced only with the program they received. They could come to the university computer lab which was reserved for this project and practiced or they could do it on-line at home if the lab time did not match their schedule. The researcher controlled their practice by checking results from the server that provided information about dates, times in, times out, number of trials, and sequence of the choices chosen by the students and met the subjects after their regular class every week. The subjects practiced with the program every other day in a linear sequence, starting form Lesson 1, Lesson 2, Test 1, Lesson 3, Lesson 4, Test 2, Lesson 5, Lesson 6, and then Test 3. The students were allowed to practice with each CALL module only once and were not allowed to repeat the materials that they had already practiced in order to control equal amount of exposure to the programs. The students completed all nine

modules within 3 weeks. If the subjects failed to follow the schedule on their checklist, the researcher would talk to them in order that they could catch up with the others.

Week 6: The researcher checked the information from the database to make sure that they all completed the program. The posttest was conducted. Also, the opinionnaire and the self-rating scales were distributed to the students.

Week 12: After the posttest, the subjects stopped practicing with the program. They returned the CD and the checklist, so they did not get any more exposure to the program. Still, they went to their regular English class. The delayed test was administered in week 12.

3.6 Data Analysis

The data were analyzed by a number of statistical analyses to answer the research questions.

1. To answer research questions 1-3, the following analyses were used:

- Means, standard deviations, and ranges of scores were employed to give a general picture of the subjects in the pretest, posttest and delayed test.

- Two-way analysis of variance (ANOVA) was used to identify the main effects and the interaction effect as posed in Questions 1, 2, and 3.

- Percentages, one-way ANOVA, and Scheffe Test were used to explain the differences elicited by Question 2.

2. To answer research question 4 concerning the retention of the two error treatments, two-way repeated measures ANOVA were used to compare the mean scores that the self-correction and the overt correction groups produced from the pretest, posttest and delayed test.

3. To answer research question 5 concerning the students' opinions and perceptions, the following statistical analyses were used:

- Frequencies and means were calculated to provide a general picture of the responses from the opinionnaire.

- One-sample t-test was applied to identify if the mean score of the SC and that of the OC group could pass the criteria set at 3.50 points.

- Chi-square test (Fisher's exact test) was applied to identify whether there was a significant difference in the proportions of the answers from the opinionnaire by students in the self-correction group and by those in the overt correction group.

- Independent-samples t-test and dependent-samples t-test were used to analyze data from the self-rating scales to identify the differences between the OC and the SC

groups and to identify the differences between pre-treatment and post-treatment scales respectively.

Chapter summary

This chapter describes the research methodology of the present study. The population consisted of 908 first-year undergraduate students of Huachiew Chalermprakiet University who took the GE 1063 course during the second semester, academic year 2006. Samples of 210 were reached by cluster random sampling. The scores were matched and then the samples were randomly assigned to the two study groups–self-correction or overt correction. Students' language abilities – high, moderate, and low– served as a moderator variable.

The experimental materials were the two CALL programs representing the two treatments. Each program consisted of 9 modules–6 lessons and 3 tests. The only one difference between the two programs was how the program dealt with errors made by the students. The overt correction program would indicate that the answer was wrong, give a full explanation, and reveal the correct answer. The self-correction program would not disclose the correct answer, but it would encourage the students to get to the correct answer by themselves with the help of the program in providing 'scaffolding' feedback.

There were 3 research instruments employed in this study. The first one was the test of English tenses usage (TETU) that served as a pretest, a posttest, and a delayed test. The second instrument was the self-rating scales on which the students were asked to rate their knowledge about the seven tenses before and after they practiced with the program. The last instrument was the opinionnaire that aimed at exploring the students' opinions towards the program they were using. All of the instruments were validated by experts and the test was also piloted with 129 students. The reliabilities of the three instruments were .90, .90, and .85 respectively.

For data analysis, descriptive statistics, two-way ANOVA, repeated measures ANOVA, one-way ANOVA, t-test, and Chi-square test were used. The results and findings for each research question are presented in Chapter IV.