CHAPTER IV
METHODOLOGY

4.0.0: Introduction 89
4.1.0: Development of Computer Assisted Language Learning package 89
4.1.1: Design and development of the CALL package 90
4.2.0: Performance assessments 92
4.2.1: Constructing a performance test 93
4.2.2: Development of Computer Managed Testing Software 94
4.3.0: Evaluation of CALL Package and Computer Managed Testing Software 96
4.4.0: Examination system and evaluation procedure 100
4.4.1: Researcher made tests 101
4.4.2: Development of Criterion Referenced Test 102
4.4.3: Validity of the test 108
4.4.4: Reliability of the test 110
4.4.5: Validity and the Reliability estimation 111
4.5.0: Pilot study 112
4.6.0: Research Design 112
4.6.1: Validity of the research design 114
4.6.2: Threats to Internal Validity 115
4.6.3: Dealing with Threats to Internal Validity 116
4.6.4: Controlling situational difference 117
4.6.5: Threats to external validity 118
4.6.6: Dealing with threats to external validity 118
4.7.0: Variables considered in the study 119
4.8.0: Tools used 120
4.9.0: Population 120
4.9.1: Need for sample collection 120
4.9.2: Sample of the present study 121
4.10.0: Experimentation 122
4.11.0: Statistical techniques used 123
4.12.0: Using Technology to Analyze Data 125
CHAPTER IV

METHODOLOGY

4.0.0: INTRODUCTION

This chapter showed how the study was carried out to observe the hypothesized relationships. A description of the population of interest, sampling procedure adopted, Development and Validation of Computer Assisted Language Learning package, Development of Criterion Referenced Test, Finding Validity and Reliability of the test, Development and Validation of computer Managed Testing Software, Research Design, Validity of research design and a description of the procedures of the present study were revealed.

4.1.0: DEVELOPMENT OF COMPUTER ASSISTED LANGUAGE LEARNING PACKAGE

(How the ideas have been acquired for developing CALL)

After deciding to prepare software to develop listening and reading skill, different readymade software packages were scrutinized carefully. Mostly all the software packages focused on introducing new words used in a familiar situation and all were beautifully animated with fun. Students should select the particular word relevant to the picture highlighted or activated. But there was no audio support. An important pre-requisite to develop listening
skill is presenting the word with proper pronunciation and pictorial representation. English Curricular CDs supplied to the CAL centers in the elementary schools were also scrutinized. Those CDs were not specific and focused on the instructional objectives selected for this study.

While searching Internet, various free trial versions of software packages were obtained and those were examined closely. Among those, one software package meant for learning German language has salient features. Having all these ideas in mind, the investigator consulted the teachers who are handling first standard English regarding their needs and comfort. The teachers expressed their view that the software would be useful when it was related to syllabus and text book oriented. After perusing the syllabus and text book, it was found that the cards utilized in the ABL methodology bearing the logos (symbols) such as Cycle rickshaw, Motor cycle and Car for standard I syllabus aim to help the students in acquiring the following competency, “Listens to and repeats the English words beginning with each letter of the alphabet” are relevant to the instructional objectives. Hence it was decided to use that particular portion as the basic skeleton for the software package.

4.1.1: DESIGN AND DEVELOPMENT OF THE CALL PACKAGE

While introducing new words in the text book, five familiar words have been given for each alphabet. Totally 130 words (5*26) have been introduced using picture cards and word cards. It was recommended by the practicing teachers and educationalists that let the same words given in the text book be used in the software package. Thus the theme was selected and
ready for the use. Then the question of how to introduce those words focusing on the purpose of listening skill development has been arisen. Again experts, practicing teachers and educationalists have been consulted.

While visiting a rural and remote elementary school, it was amazing to notice that all the students recited English rhymes clearly with correct pronunciation. Mostly all the students were socially deprived and learners of first generation. Out of curiosity, the investigator asked the teacher to reveal the technique he followed to attain that achievement. He showed the English note books of his students, in which, his students had written English words in Tamil (students' mother tongue) using transliteration technique.

The investigator consulted an eminent English professor who is working at CIEFL (Central Institute of English and Foreign Languages) Hyderabad about how far mother tongue or transliteration technique could be used for research purpose. After gaining more suggestions, an innovative strategy has been used in the transliteration technique. As already mentioned, English is a highly un-phonetic language whereas students' mother tongue TAMIL is a phonetic language. Apart from this, there are some new sounds which do not exist in Tamil language. Moreover in Tamil, for one letter, there are many sounds. For example, “K” letter is used for “k” as well as “g”. “P” letter is used for “p” and “b”. “T” letter is used for “t” and “d”. Hence it was decided to use two dots below the alphabet of Tamil language to have voiced sound while using transliteration technique.
Again after consulting with software developers, practicing teachers and educationalists, a software package has been developed having four stages. In the first stage, a word card was shown. This card was meant for learning the correct pronunciation of the word by looking at the picture and listening to the audio file. In that card, there was a picture with a relevant word in English as well as written in Tamil language using transliteration technique. A speaker symbol was there. On clicking the speaker symbol, the word given was pronounced. Repetition was possible. After getting enough practice, students clicked ‘NEXT’ button to go to the next word. There were five such word cards for each alphabet. Then the menu “List” would be clicked for selecting the next or preferred alphabet. This first stage was meant for introducing new words using drill and practice method and thus served as Computer Assisted Language Learning package. The copies of the first page mentioned as “Know the word” and the page “Change List” to select the desired alphabet were given in the Appendix No 1. The underlying principle of the remaining three stages was, assessing recognition, listening and reading skills and thus utilized as Computer Managed Testing Software.

4.2.0: PERFORMANCE ASSESSMENTS

The most common achievement tests are paper-and-pencil tests measuring cognitive objectives. Everyone is familiar with this format that requires the person to compose answers or choose responses on a printed sheet. In some cases, however, a researcher may want to measure performance-what an individual can do rather than what he or she knows. A performance test is a technique in which a researcher directly observes and
assesses an individual's performance of a certain task and judges the finished product of that performance. The performance is judged against established criteria. Performance assessment is important in areas such as art, music, public speaking and industrial training. Portfolios that contain a collection of student work such as poetry, essays, sketches, musical compositions, audiotapes of speeches and even mathematics worksheets are a popular tool in performance assessment. (Donald Ary 2006).

Stiggins (2001), an early advocate of performance assessment, defines it in terms of four components.

- a reason for the assessment
- a particular performance to be evaluated
- exercises that elicit the performance, and
- systematic rating procedures

4.2.1: CONSTRUCTING A PERFORMANCE TEST

When a performance assessment is desired, the researcher should follow three basic steps in constructing the test.

1. Begin with a clear statement of the objectives or what individuals should do on the test and the conditions under which the task will be performed.

2. A set of test specifications listing the critical dimensions to be assessed will lead to a more comprehensive coverage of the domain.

3. State whether there will be time limits, whether reference books will be available and so on.
Provide a problem or an exercise that gives students an opportunity to perform – either a simulation or an actual task. All individuals should be asked to perform the same task. Develop an instrument (check list, rating scale, or something similar) that lists the relevant criteria to use in evaluating the performance and/or the product. To maximize objectivity, the instrument should ensure that the same criteria are used for each individual’s performance or product.

Performance tests provide a way to measure abilities and skills that cannot be measured by paper-and-pencil tests. They are, however, time intensive and thus more expensive to administer and score. These ideas were contemplated genuinely while developing Computer Managed Testing Software.

4.2.2: DEVELOPMENT OF COMPUTER MANAGED TESTING SOFTWARE

The investigator developed a software program using .Net to administer the test on academic achievement in English. As the multiple choice items are the widely used form of objective type test items, the investigator used the same form in the test. “English listening skills learning for beginners include English listening comprehension exercises for basic listening skills such as the alphabet, words, numbers and simple dialogues". Based on this idea, the test items were developed.

In “Know the picture” stage, a plain picture card was shown without the support of audio file and printed word. After recognizing the word by looking at the picture, students clicked the correct one from the options given.
If he/she clicked correctly, automatically one mark has been awarded and was shown in the specified box immediately. After that they clicked the next button to see the next picture. For further correct answers, cumulative total was shown. Totally five marks were awarded for recognizing an alphabet.

Next “Know the Accent” stage was displayed. A speaker symbol with five options was shown in the card. Students clicked the speaker symbol to hear the pronunciation of the word. After listening to the audio file, they recognized the word and clicked the correct option. Again same scoring procedure was followed. In this stage, five marks were awarded for the listening comprehension of an alphabet.

Finally, “say the word” card meant for assessing the reading skill was shown. On clicking the correct option, Computer itself awarded marks while assessing recognition and listening skills. Since Automatic Speech Recognition feature was not used, teacher assistance was required to assess the reading skill. When the student read the word correctly which was shown in the card, the teacher clicked the right symbol for the correct pronunciation and the wrong symbol for the incorrect pronunciation. Then the same scoring procedure was followed. To sum up, the computer managed testing software program was developed in such a way that the scores of the individual student under the classification of recognition, listening and reading skills for every alphabet were shown individually on the screen and finally the total marks awarded out of fifteen was shown.
After completing all the three stages for all the 26 alphabets, the final total score awarded out of 390 marks was taken into account for analysis. The copies of the second stage (Know the picture), third stage (Know the Accent), fourth stage (say the word) and the final score box were given in the Appendix No 1. Out of 520 cards used, the copies of the 130 (26 alphabets * 5) “First Stage” picture cards were shown in the Appendix No 2.

4.3.0: EVALUATION OF CALL PACKAGE AND COMPUTER MANAGED TESTING SOFTWARE

Evaluating CALL software is not a straightforward job. It's quite different from evaluating a book. A good deal of the contents of a CALL software package will not be immediately visible and will only appear if the user follows a particular route. In addition, there are factors such as screen design, user-friendliness, nature of the interaction, etc to take into account. The key issue, as Chris Jones pointed out as long ago as 1986, is: "It's not so much the program: more what you do with it" (Jones C. 1986).

Whatever kind of technological programs educators decide to develop, the primary consideration should be the purpose of using the technology. Jordan and Follman (1993) outline a number of characteristics of good technology programs. They should

- emphasize cooperative learning models, allowing heterogeneous groupings of students to work together collaboratively;
- support interactions between students and teachers rather than use computers as "teaching machines" to supplant the teacher;
- be adaptable to a variety of learning styles.

96
In addition, as Cole and Griffin (1987) point out, it is essential that programs be designed to ensure equitable and substantial access for all students, and function as an integral part of a well-planned pedagogy. Educators planning to introduce technology programs into their schools must also consider a number of practical matters in order for these programs to be successful. As Fulton puts it, technology requires that schools be willing to make substantial investments in time, resources, and support (1993, p. 3). On the most obvious level, for example, someone in the school must know how to install the equipment and keep it working properly. Further, as a 1990 study by the Center for Technology in Education (cited in Fulton, 1993) found, even when teachers are believing about the appropriateness of educational technology and are willing to learn, they can take as much as five or six years to become sufficiently comfortable with computers to able to use them effectively in their classrooms.

ASPECTS OF EVALUATION

Aspects which should be evaluated to determine the quality of the software are:

- Design of the programme
- Content
- Pedagogical approach
- Language
- Illustrations
All the above aspects are equally important and their level should be judged to be at least very satisfactory if the final form is to be approved. Imperfections or an inferior in one aspect will inevitably have negative consequence on the other. For example poor quality illustrations, which are blurred or inaccurate, will distract the user when he tries to relate them to explanations and information contained in the body of the program. Presentation and composition of the text should also be satisfactory.

These aspects of evaluation are in general, all valid for CAI packages also. However, differences do exist depending upon the discipline under consideration and the year or level of instruction. The same aspects will also differ for different subjects. Before applying evaluation criteria, therefore the CAI material must be suited in relation to the aspects evaluated and its final destination.

For evaluating CALL software, the following section of the ICT4LT website was referred. http://www.ict4lt.org Section 3.8, Module 1.4.

**Software Evaluation Form**

<table>
<thead>
<tr>
<th>Title of software package / program: Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Is the level of language that the program offers clearly indicated?</td>
</tr>
<tr>
<td>2. Is it easy to start the program?</td>
</tr>
<tr>
<td>3. Is the user interface easy to understand?</td>
</tr>
<tr>
<td>4. Is it easy to navigate through the program?</td>
</tr>
<tr>
<td>5. Are icons that are used to assist navigation (e.g. back to the homepage, exit) clear and intelligible?</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td><strong>6. Is it always clear to the learner which point s/he has reached in the program?</strong></td>
</tr>
<tr>
<td><strong>7. Does the program include scoring?</strong></td>
</tr>
<tr>
<td><strong>8. If a scoring system is used, does it make sense?</strong></td>
</tr>
<tr>
<td><strong>9. If a scoring system is used, does it encourage the learner?</strong></td>
</tr>
<tr>
<td><strong>10. Is the learner offered useful feedback if s/he gets something wrong?</strong></td>
</tr>
<tr>
<td><strong>11. If the learner gets something right purely by chance, can s/he seek an explanation in order to find out why the answer is right?</strong></td>
</tr>
<tr>
<td><strong>12. Can the learner seek help, e.g. on grammar, vocabulary, pronunciation, cultural content?</strong></td>
</tr>
<tr>
<td><strong>13. Does the program branch to remedial routines?</strong></td>
</tr>
<tr>
<td><strong>14. Can the learner easily quit something that is beyond his/her ability?</strong></td>
</tr>
<tr>
<td><strong>15. Are the grammar and vocabulary used in the program accurate?</strong></td>
</tr>
<tr>
<td><strong>16. Does the program offer cultural insights?</strong></td>
</tr>
<tr>
<td><strong>17. If the program includes pictures, are they (a) relevant, (b) an aid to understanding?</strong></td>
</tr>
<tr>
<td><strong>18. If the program includes sound recordings, are they of an adequate quality?</strong></td>
</tr>
<tr>
<td><strong>19. If the program includes sound recordings, are they (a) relevant, (b) an aid to understanding?</strong></td>
</tr>
<tr>
<td><strong>20. Is the program relevant to your national / Regional / departmental programme of study?</strong></td>
</tr>
<tr>
<td><strong>21. Does the software offer anything extra that cannot be done in more traditional ways, e.g. with pencil and paper</strong></td>
</tr>
<tr>
<td>Question</td>
</tr>
<tr>
<td>-------------------------------------------------------------------------</td>
</tr>
<tr>
<td>22. Does the program make use of Speech Engine so as to listen to native speaker’s voice?</td>
</tr>
<tr>
<td>23. Does the program make use of Speech Recognition so as to record his/her own voice and hear the playback clearly?</td>
</tr>
</tbody>
</table>

Using the above “Software Evaluation Form” developed by ICT4LT Project 2009 under a Creative Commons Attribution-Noncommercial-No Derivative Works 2.0 UK: England & Wales License, the Computer Assisted Language Learning package and the Computer Managed Testing Software were evaluated by 30 experts comprising practicing teachers, educationists, education technologists and software developers.

The experts were supplied with the evaluation form and requested to register their opinion. From the analysis of the responses made by the experts, it was found that out of 23 questions except 11 and 23; all other questions were answered positively. 91.3% of the criteria were fulfilled. Hence it was proved that the software package prepared was appropriate.

**4.4.0: EXAMINATION SYSTEM AND EVALUATION PROCEDURE**

In India, generally the examination in English puts a lot of premium on rote learning rather than on language mastery. There is no examination in spoken English. At the time of examination, much importance is given to written English.

The evaluation procedure is purely based on essay type of questions. Hence it could not measure the entire development of the student in English. The marks awarded to the students do not entirely indicate the level of
achievement of the students. The traditional evaluation process measures only the knowledge aspect of the student and do not measure the skill.

Hence the examination system needs to be modernized. In this respect the report of English Review Committee (1965) suggests that “The evaluation in English should be in line with the new technique of teaching-learning. Stress should be laid on the testing of oral work and pronunciation.”

4.4.1: RESEARCHER MADE TESTS

Selecting appropriate and useful measuring instruments is critical to the success of any research study. Researcher must select or develop scales and instruments that can measure complex constructs such as intelligence, achievement, personality, motivation, attitudes and so on. Achievement tests are widely used in educational research. They measure mastery and proficiency in different areas of knowledge by presenting subjects with a standard set of questions involving completion of cognitive tasks. Achievement tests are generally classified as either standardized or researcher made. In selecting an achievement test, researchers must be careful to choose one that is reliable and is appropriate (valid) for measuring the aspect of achievement in which they are interested.

To select a measuring instrument, the investigator looked at the research that has been published on her problem to see what other researchers have used to measure the construct of interest – achievement of the students of standard I in acquiring listening and reading skills in English. The investigator looked at the Mental Measurements Yearbooks and Tests in print
and Educational Testing Service (ETS) Test Collection for identifying published instruments and for up-to-date test information. But the investigator could not find a previously developed apt instrument and hence decided to develop her test.

When using standardized tests of achievement is not considered suitable for the specific objectives of a research study, research workers may construct their own test. It is much better to construct our own test than to use an inappropriate standardized one just because it is available.

4.4.2: DEVELOPMENT OF CRITERION REFERENCED TEST

A criterion-referenced test is one that provides for translating test scores into a statement about the behavior to be expected of a person with that score or their relationship to a specified subject matter. Most tests and quizzes written by school teachers are criterion-referenced tests. The objective is simply to see whether or not the student has learned the material. A common misunderstanding regarding the term is the meaning of criterion. Many, if not most, criterion-referenced tests involve a cut score, where the examinee passes if their score exceeds the cut score and fails if it does not (often called a mastery test). The criterion is not the cut score; the criterion is the domain of subject matter that the test is designed to assess. For example, the criterion may be "Students should be able to correctly add two single-digit numbers," and the cut score may be that students should correctly answer a minimum of 80% of the questions to pass.
The criterion-referenced interpretation of a test score identifies the relationship to the subject matter. In the case of a mastery test, this does mean identifying whether the examinee has "mastered" a specified level of the subject matter by comparing their score to the cut score. However, not all criterion-referenced tests have a cut score, and the score can simply refer to a person's standing on the subject domain. (From Wikipedia, the free encyclopedia)

Criterion-referenced Language Testing, authored by James Dean Brown and Thom Hudson (2002), asserts how criterion-referenced testing (CRT) can provide realistic and useful test development tools that will assist language teachers and language curriculum developers in their respective jobs. In fact, over the past decades, CRT, which provides information about an individual's mastery of a given criterion domain or ability level, has become an emerging issue in language assessment, especially in language achievement tests.

Criterion referenced tests enable researchers to describe what a specific individual can do without reference to the performance of others. Criterion referenced tests are used to determine an individual's status with respect to a well defined set of content objectives. Performance is reported in terms of the level of mastery of some defined content or skill domain. Typically, the level of mastery is indicated by the percentage of items answered correctly.

CRTs give detailed information about how well a student has performed on each of the educational goals or outcomes included on that test.
For instance, "... a CRT score might describe which arithmetic operations a student can perform or the level of reading difficulty he or she can comprehend" (U.S. Congress, OTA, 1992, p. 170). As long as the content of the test matches the content that is considered important to learn, the CRT gives the student, the teacher, and the parent more information about how much of the valued content has been learned than a norm-referenced test (NRT).

In norm-referenced tests, items are selected that will yield a wide range of scores. A researcher must be concerned with the difficulty of the items and the power of the items to discriminate among individuals. In criterion referenced tests, items are selected solely on the basis of how well they measure a specific set of instructional objectives. They may be easy or difficult, depending on what is being measured. The major concern is to have a representative sample of items measuring the stated objectives so that individual performance can be described directly in terms of the specific knowledge and skills that these people are able to achieve. (Donald Ary 2006)

Rules to be observed while developing the criterion referenced test:

1. Identification of Domains

A domain is described as a chunk of course content called out of a particular topic or unit of learning. This chunk of syllabus may vary from small scale piece of information to the largest segment of knowledge in that area. This segment of content which may vary from smallest piece of
information to that of the biggest segment of knowledge may be considered as a domain from the point of view of testing.

2. Strategy for Test Construction

Once the basis or the framework is ready, two strategies are possible for test construction. One may select number of domains and test for basic language skills Listening, Speaking, Reading and Writing together. Otherwise we may test only the key idea presenting the core intending learning outcomes.

3. Selection of specific Objectives

General objectives which are supposed to have been identified earliest on the basis of importance, ease of production and contents scorability, skills etc do not enable us to develop test which can be interpreted in terms of predetermined criteria unless we formulate specific objectives for every domain.

4. Developing Achievement Continuum

As stated already the domain description envisages identification of not only the content hierarchy or sequence but also the hierarchy of learning outcomes. The type of judgment in terms of expected or intended performance as implied in a behaviorally stated specific objective of the domain had to be thought by the investigator necessarily.

5. Identification of the subject area

The first step in development of criterion referenced test is the decision about the subject matter area to be worked out. For example one may take up
English, Mathematics, Physics etc, depending upon the need of the area and the resources one can utilize for one or more subject areas.

6. Selection of the unit topic

After the decision is taken about area, the next step is to select the unit on which test is to be developed. Depending upon the need, one or more than one section or chunks of content may be taken up for construction.

7. Description of domain of testing

Since domain refers to a particular segment of the content, one may examine the topic and define it into various segments which can be developed into well defined separate domains. Each domain can then be analyzed in terms of facts, concepts, principles, process etc. that may be arranged in order of their increasing complexity. It should reflect clearly the nature and scope of the content specification in sequential, hierarchical or developmental order.

8. Specification of Domain objectives

Having decided about the content elements of a domain selected, the next task is to formulate the instructional objectives or expected learning outcomes which may be categorized in terms of Listening, Speaking, Reading and Writing. There should be precise statement of objectives and that the pupil's performance could be interpreted clearly in terms of the outcomes of intended learning. Sample prototype items framed may accompany each specific objective for more clarification.

9. External review of steps
The tasks identified should be reviewed by those who are not involved in identifying the domain and its descriptions in terms of content elements and the specific objectives. However, a teacher who teaches English to the beginners also involved in this task so that he/she may be able to clarify doubts if any raised by the external reviewers. Major purpose of this review is to sharpen domain description and the specific objectives in order to make them more realistic and functional.

10. External review of the test

Prior to the field trial the test is to be re-examined by the practicing teachers. The main purpose is to detect some content flaws, if any, and check for congruence of things with the domain description. It is to be done preferably by those same people who do external review of steps.

11. Field trial of test

At this stage the test is to be tried out on a limited number of students, say 5 to 10 to get a fix on the elements in the instruction which might be proved drastically wrong. It is better that the subject teachers other than those involved in the development of the test, administer the test. However, one team member of the developers may accompany them in order to meet queries if any, relating to the content or procedure that might arise during try out of the test. This is possible only if the member who is familiar with the development of the test is associated in the field try out.

12. Internal review
After the field trial of the test, internal review would provide a last look at the test which would depend on the changes, if any that have been made as a result of the review or the field try out. The purpose of this review by the investigator is to certify the final format of the test and pass it for use.

13. Final form of the test

At this stage, the final form of the test is ready for use and administration. A sign off sheet used to accompany the various tests as they move from above steps. Each test is kept in a folder to which the signed sheet may be attached. This enables the developer to keep track of the given test. To monitor the progress of the test as it goes through the various stages of development, a master progress sheet is used.

14. Using the test in the classroom

The test is to be administered to test the domains which are covered by the test. The domains being tested are arranged according to the needs of the teacher and administered one after the other in sequential order. Students' responses can be recorded and tabulated in accordance with the scheme of analysis which is mostly in terms of specified domain objectives.

15. Finding validity and reliability of test

Since the data are available on the test, the reliability and validity of the test can be found out using various techniques.

Almost all the above said principles were strictly observed while developing the criterion referenced tests in the content areas selected for instruction during experimentation.
4.4.3: VALIDITY OF THE TEST

Validity is the most important consideration in developing and evaluating measuring instruments. Historically, Validity was defined as the extent to which an instrument measured what it claimed to measure. The focus of recent views of validity is not on the instrument itself but on the interpretation and meaning of scores derived from the instrument. The most recent Standards for Educational and Psychological Testing (1999), prepared by the American Educational Research Association (AERA), National Council on Measurement in Education (NCME), and the American Psychological Association (APA) defines validity as “the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests”.

The Standards for Educational and Psychological Testing lists three categories of evidence that can be used to establish the validity of score based interpretations: evidence based on content, evidence based on relations to a criterion and construct related evidence of validity.

Evidence based on test content involves the test’s content and its relationship to the construct it is intended to measure. The Standards define content-related evidence as “the degree to which the sample of items, tasks or questions on a test are representative of some defined universe or domain of content”. This type of evidence is especially important in evaluating achievement tests. Brennan (2001) writes that “In educational achievement tests, content-related validity evidence is absolutely essential. If the content can’t be defended little else matters” (p.12).
Bearing in mind the above points, the investigator first experienced what kinds of skills and objectives are covered in first Standard English classes throughout the system. After examining text books and talking with teachers, the investigator prepared an outline of the topics, skills and performances that make up first Standard English (Content Domain) in the system, along with an indication of the emphasis given to each. Using the outline as a guide, the investigator had written a collection of test items that covered the content domain and proper emphasis was given to the instructional objectives. The test items written were a representative sample of the total domain of knowledge and skills included in the first Standard English syllabus.

Teachers, Educationists and subject experts were asked to look at the test content to judge the appropriateness and representation of the items making up the test. The instructional objective for the present study was "Listens to and repeats the English words beginning with each letter of the alphabet". To accomplish this, five words for each alphabet (5*26) totally 130 words have been given in the syllabus. Since English is a highly un- phonetic language, all the 130 words were taken into consideration.

4.4.4: RELIABILITY OF THE TEST

Reliability indicates how consistently a test measures whatever it does measure. Specifically reliability refers to the extent to which an individual scores nearly the same in repeated measurements.
Donald Ary (2006) elaborated the factors affecting reliability of a test as follows:

<table>
<thead>
<tr>
<th>Factor</th>
<th>Potential effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Length of the test</td>
<td>The longer the test, the greater the reliability</td>
</tr>
<tr>
<td>2. Heterogeneity of group</td>
<td>The more heterogeneous the group, the greater the reliability</td>
</tr>
<tr>
<td>3. Ability level of group</td>
<td>Test that is too easy or too difficult for group results in lower reliability</td>
</tr>
<tr>
<td>4. Techniques used to estimate reliability</td>
<td>Test-retest and split-half give higher estimates. Equivalent forms give lower estimates.</td>
</tr>
<tr>
<td>5. Nature of the variable</td>
<td>Tests of variables those are easier to measure yield higher reliability estimates.</td>
</tr>
<tr>
<td>6. Objectivity of scoring</td>
<td>The more objective the scoring, the greater the reliability.</td>
</tr>
</tbody>
</table>

The investigator considered all these above points carefully while developing the test. The length of the test is extensive and all questions were objective type only. Furthermore attention was given so that the test was neither too easy nor too difficult.

**4.4.5: VALIDITY AND THE RELIABILITY ESTIMATION**

To establish the validity of the test, the investigator attempted to find out the correlation co-efficient between the achievement scores in English obtained from this test and the scores in English in the school examination by product moment correlation co-efficient method. The value of ‘r’ is found to
be 0.886 and is significant at 0.01 level. Hence it is concluded that test has high validity.

To establish the reliability of the test, the investigator adopted the split-half method. "The test is split into two equivalent halves by pooling the odd numbered items for one score and the even numbered items for another score. This usually makes the two scores obtained from a single test reasonably equivalent. In this way two scores for each student are obtained, one on odd numbered items and the second on the even numbered items. The correlation between the results of the two halves is determined and from these, the reliability of the whole test is calculated by applying the Spearman-Brown Prophecy formula".

The investigator adopted the same procedure for establishing the reliability of the test by using the score of the post test. The reliability of the whole test was found to be 0.895 and was significant at 0.01 level. Hence it was concluded that the test was highly reliable.

4.5.0: PILOT STUDY

The pilot study was conducted to know how good the software was to accomplish the defined objectives. Municipal middle school which is a CALL center in Gobichettipalayam block and has all the computer facilities was selected for this purpose. The investigator administered a draft of the test to a small group (10 students) who has not participated in the experimentation. The pilot study helped the researcher to decide whether the study was practicable and whether it was sensible to continue. It provided an
opportunity to assess the appropriateness of the research methodology. An analysis of the results facilitated the investigator to check on the test’s validity and reliability and to detect any ambiguities or other problems before utilizing the test. Based on the analysis, refinement was made in the test.

4.6.0: RESEARCH DESIGN

An experiment is a scientific investigation in which the researcher manipulates one or more independent variables, controls any other relevant variables and observes the effect of the manipulations on the dependent variables. An experimenter deliberately and systematically introduces change and then observes the consequences of that change. The goal of experimental research is to determine whether a causal relationship exists between two or more variables. In the 1890s the experimental method was first used to study an educational problem. Rice’s (1897) investigation of spelling achievement in the schools marks the first attempt at educational field experimentation. Thorndike (1924) and other early investigators extended the experimental method to education.

Educational Research is the application of the scientific approach to the study of educational problems. Educational research often does not allow for the kind of randomization noted for true experiments. In many situations in educational research, however, it is not possible to randomly assign subjects to treatment groups. Neither the school system nor the parents would want a researcher to decide to which classrooms students were assigned. In this case, researchers turn to quasi-experiments in which random assignment to treatment groups is not used. Quasi-experimental designs are similar to
randomized experimental designs in that they involve manipulation of an independent variable but differ in that subjects are not randomly assigned. Campbell and Stanley (1966) stated, however, that quasi-experimental studies are "well worth employing where more efficient probes are unavailable" (p. 205). These designs permit the researcher to reach reasonable conclusions even though full control is not possible. Quasi experimental method was found to be the most appropriate method for testing the spelt out hypotheses in the present study.

**Use of control groups**

This is the most used and established way of exercising control over significant variables. The principle is simple. First, it is necessary to identify two groups of people for the experiment. The two groups should be similar in terms of their composition. The start point for the experiment then is two matched groups or samples. The experiment involves introducing a factor to the experimental group and leaving the other group with no artificially induced changes. Having added the new factor, the experimenter can then look at the two groups again with the belief that any difference between the groups can be attributed to the factor which was artificially induced.

Certain changes over time are likely to happen irrespective of whether an experiment had taken place. Time moves on for all of us. But, if the two groups were as identical as possible at the outset, any change of this kind in one group will also occur in the other group. So, instead of measuring the change from time 1 to time 2, we measure the difference between the control
group and the experimental group at the end of the experiment at time 2. Any differences we observe can be logically deduced to come from the factor which was artificially induced in the experiment. - Martyn Denscombe (1999)

4.6.1: VALIDITY OF THE RESEARCH DESIGN

Researchers have to be concerned with the correctness of the inferences they draw about a causal relationship between the variables of the study. A very significant contribution to the validity of experiments was made by Campbell and Stanley (1963), who suggested two general categories of validity: internal validity and external validity. Internal validity: The validity of the inferences about whether the effect of variable A (treatment) on variable B (the outcome) reflects a causal relationship.

External validity: The validity of the inferences about whether the cause-effect relationship holds up with other subjects, settings and measurements.

4.6.2: THREATS TO INTERNAL VALIDITY

1. History - Unrelated events that occur between pre and post tests affect the dependent variable

2. Maturation - Changes occur within the participants just as a function of time.

3. Testing - Exposure to prior test affects posttest.

4. Instrumentation - Unreliability or a change in the measuring instrument affects result.

5. Regression - Extremely high or low scorers on a pretest regress toward mean on a posttest.
6. Selection - Because of selection methods, subjects in the comparison groups are not equivalent prior to study.

7. Mortality - A differential loss of participants from the groups affects dependent variable.

8. Selection-maturation interaction - Subjects with different maturation rates are selected into treatment groups.

9. Experimenter effect - Unintentional bias or behavior of experimenter affects result.

10. Subject effect - Attitudes developed during the study affect performance on dependent variable.

11. Diffusion - Participants in experimental group communicate information about treatment to control group which may affect the latter's performance.

### 4.6.3: DEALING WITH THREATS TO INTERNAL VALIDITY

An experiment should be designed to avoid or at least minimize the effect of threats to internal validity. Reichardt (2000) states that it is better to speak of taking account of threats to validity than ruling out threats because the latter implies a finality that can rarely be achieved in practice. The researcher's first efforts must be directed toward controlling for any relevant preexisting differences between subjects used in an experiment. Only in this way the investigator was confident that any post experimental differences could be attributed to the experimental treatment rather than to pre existing subject differences.
Six basic procedures are commonly used to increase equivalence among the groups that are to be exposed to the various experimental situations: 1. random assignment, 2. randomized matching, 3. homogeneous selection, 4. building variables into the design, 5. analysis of covariance, and 6. use of subjects as their own controls. Homogeneous selection procedure was followed to increase the equivalence among the groups. To make groups reasonably comparable on an extraneous variable is to select samples that are as homogeneous as possible on that variable. "As random selection of subjects was not possible, a variety of demographic and aptitude comparisons were made to determine if the control and experimental groups were similar."

4.6.4: CONTROLLING SITUATIONAL DIFFERENCE

Extraneous variables may operate in the experimental setting to create situational differences that can threaten internal validity. These uncontrolled extraneous variables whose effects may mistakenly be attributed to the independent variable are called confounding variables. Holding extraneous variables constant means that all subjects in the various groups are treated exactly alike except for their exposure to the independent variable.

In the present study, experimental groups' students were sent off to the computer room to utilize the CALL package, leaving the control group students in the classroom. In such situation, the students in the control group may be resentful or feel rejected or inferior to the others. The use of a placebo is called a single-blind experiment in which the subjects are unaware of the treatment condition they are in, although the researcher does know.
For this reason, in the present study the subjects of the control group had been shown some other English curricular CD which was not related to the construct of interest of the present study. Steps had been taken to ensure that the control subjects also had an opportunity of employing CALL package and all the three groups were unacquainted that they were taking part in an experiment. This precaution was necessary in order to control subject effects. Since the sample was students of standard one (5+ aged), actually they did not sense anxiety as well as fatigue throughout the experimentation.

In the present study, experimental groups and the control group had the same number of subjects. It was noted that the teacher efficacy and enthusiasm affected the outcome of the learning experiment during the pilot study. To control the teacher variable, since it was not possible to have the same teacher for the treatments, available teachers were randomly assigned to the three groups. In the present study the investigator followed the following procedures: same instructions, same computer requirements, same tests etc. All the three groups were given instruction at the same time of the day and in the same type of a room. In this way the situational variables were randomized.

4.6.5: THREATS TO EXTERNAL VALIDITY

1. Selection - An effect found with certain kinds of subjects might not apply if other kinds of subjects were used.

2. Setting- treatment interaction - An effect found in one kind of setting may not hold if other kinds of settings were used.
3. Pretest- Pretest may sensitize subjects to treatment to produce an effect not generalizable to an unpretested population

4. Subject effects - Subjects’ attitudes developed during study may affect the generalizability of the results.

5. Experimenter effects - Characteristics unique to a specific experimenter may limit generalizability to situations with a different experimenter

6. Novelty effect - A novel treatment may lead to excitement among subjects that causes them to respond differently than they would in a normal situation.

**4.6.6: DEALING WITH THREATS TO EXTERNAL VALIDITY**

Before assuming external validity, the investigator examined the similarities and differences between the experimental setting and the target setting with respect to subjects and treatments. The following steps had been taken to control threats to external validity. Subjects were selected randomly from the experimentally accessible population.

In the present study, the similarity of the experimentally accessible population and the target population is the age 5+ and the difference is availability of computer facilities in the schools.

Since the investigator wanted to generalize the effect, different strata like rural-urban, government-aided-unaided schools and mixed-girls schools were included and interpreted the performance of each group separately to see if the experimental treatment worked well with all groups.
Since the subjects were of the age group 5+, they were innocent and unaware of the experiment went on and hence no novelty effect.

4.7.0: VARIABLES CONSIDERED IN THE STUDY

Variables are the conditions or characteristics that the experimenter manipulates, controls, or observes. In educational research, an independent variable may be a particular teaching method, a type of teaching material. The dependent variable may be a test score in performing a task. Thus the dependent variables are the measured changes in students' performance attributable to the influence of the independent variable.

INDEPENDENT VARIABLE

Computer Assisted Language Learning software package was created and used as an instructional strategy to facilitate to acquire recognition, listening and reading skills in English.

DEPENDENT VARIABLE

Test scores obtained in performing the task designed to measure recognition, listening and reading skills in English.

4.8.0: TOOLS USED

- Syllabus based computer software package focusing on recognition, listening and reading skills prescribed for first Standard English was developed using the computer language C# (C Sharp) and the framework .Net (Dot Net) and has been evaluated by practicing teachers, educationalists and technical experts.
• Computer managed testing software was developed following the same procedure to administer pre, post and retention tests to the students of the three groups.

4.9.0: POPULATION

Children who are completed five years of age admitted to first standard in the elementary schools are considered as the students of standard I. Since the study focused on the effectiveness of CALL on recognition, listening and reading skills of students of standard I in English, the students of standard I of both sexes studying in elementary schools which have been selected as CAL (Computer Assisted Learning) centers were considered as the population. The students selected for the study belong to the academic year 2006-2007.

4.9.1: NEED FOR SAMPLE COLLECTION

Every research investigation needs a sample selection because the whole population may not be taken for the study. It would be impracticable to test or to observe each unit of the population under controlled conditions in order to arrive at principles having universal validity. Some populations are so large that their study would be expensive in terms of time, energy, money, effort and manpower. Sampling is the process by which relatively small number of individuals has been selected from the population and analyzed in order to find out something about the entire population.

4.9.2: SAMPLE OF THE PRESENT STUDY
As the population is spread over the whole of the Gobichettipalayam block, visiting all these schools and approaching all the subjects of population becomes hardly practicable. So the investigator has decided to select the sample from the population using stratified sampling procedure. The stratified sampling procedure is applicable when the population is composed of subgroups or strata of different sizes. The sample for the final test was drawn from four schools in Gobichettipalayam block.

The schools which are managed directly under the control of State Government and supervision by department of education are called Government schools. The schools which are run by private authorities but aided by Government are called Aided schools. The schools which are run by private authorities and not aided by Government are called unaided schools. These schools are subdivided on the basis of location namely rural and urban area schools. These schools are further classified as mixed schools and Girls schools.

The sample of 120 students studying at first standard in four schools in Gobichettipalayam block was selected as below.

Panchayat Union Middle School, Karattadipalayam 30
(Government, Rural, Mixed)

Panchayat Union Middle School, P.Vellalapalayam 30
(Government, Rural, Mixed)

Shri saradha Vidhyalaya School, Gobichettipalayam 30
(Aided, Urban, Mixed)
All the above mentioned students were of same age group and belong to similar socio economic status.

4.10.0: EXPERIMENTATION

Pre – Post test equated group design was found to be the most appropriate method for testing the spelt out hypotheses in the study. Three identical groups each of 40 first standard students were formed based on their pre achievement level in English. One of the groups formed as control group while the other two groups formed as experimental groups.

Software package in the selected content area in first Standard English syllabus had already been developed and evaluated. The same content was taught to all the three groups through respective instructional strategy. The students of all the three groups were given instruction in the selected content area for 26 days, an alphabet per day. Traditional method was adopted in the control group, while only CALL as instructional strategy and CALL as a support system to teachers’ class room instruction were introduced as experimental factors in the experimental groups.

Pre and post tests in the same content area were administered to all the three groups. Criterion referenced test was used for the control group and students of the experimental groups were given test through the specially developed Computer Managed Testing Software. Retention tests in the same content area were also administered to all the three groups after the annual
holidays roughly one month after the experiment was over. The scores obtained by the students of control and experimental groups on academic achievement in English were computed for pre, post and retention tests. The spelt out hypotheses were tested using appropriate statistical techniques.

4.11.0: STATISTICAL TECHNIQUES USED

"Many statisticians suggest that parametric tests be used, if possible, and that nonparametric tests be used only when parametric assumptions cannot be met. Others argue that nonparametric tests have greater merit than is often attributed to them because their validity is not based upon assumptions about the nature of the population distribution, assumptions that are so frequently ignored or violated by researchers employing parametric tests." - John W. Best and James V.Kahn (2004).

When the data are interval or ratio scaled and the samples size is large, parametric statistical procedures are appropriate. These procedures are based on the assumptions that the data in the study are drawn from populations with normal distributions and normal sampling distribution. "If a parametric test is used when the data are not parametric, then the results are likely to be inaccurate. Hence it is very important that the assumptions should be checked before deciding which statistical test is appropriate." - Andy Field (2005) - Discovering Statistics using SPSS.

"Non-parametric tests are designed for specific populations, for example, a tear group in a school or teachers of a particular subject. They are
much quicker to apply and the calculations involve are often less problematic than for parametric tests”. - Clive Opie (2004).

“Non-parametric tests are the “stock-in-trade of class room teachers”.

“Nonparametric tests are appropriate when the nature of the population distribution from which samples are drawn is not known to be normal” - John W. Best (2004).

For the compliance of Normality and Homogeneity of variances, it was decided to conduct the Kolmogorov-Smirov test of Normality and Levene's test for Homogeneity of variances. Based on the Kolmogorov-Smirov test result; it was found that the distributions were not significantly normal. Hence it was decided to apply the nonparametric tests which do not rest upon the more stringent assumption of normally distributed populations.

Of the many nonparametric tests, Kruskal-Wallis test (equivalent to one way ANOVA ‘F’ test) and Mann-Whitney test were used. Kruskal-Wallis test is a non-parametric test used when one wants to find out if three or more independent groups belong to a single population.

“The Mann-Whitney test is designed to test the significance of the difference between two populations, using random samples drawn from the same population. It is a nonparametric equivalent of the parametric t test. It may be considered a useful alternative to the t test when the parametric assumptions cannot be met and when the observations are expressed in at least ordinal scale values.” – John W. Best (2004).
4.12.0: USING TECHNOLOGY TO ANALYZE DATA

The two most common statistical programs used in the social sciences, including education, are SPSS (Statistical Package for the Social Sciences) and SAS (Statistical Analysis System). Both SPSS and SAS were initially developed in the 1960s. Both can import data from most spreadsheet software (such as Excel) and from databases (such as Access). The investigator used SPSS technological tool to describe data and calculate the relevant statistics.