CHAPTER 3

METHODOLOGY

AND

DATA COLLECTION
CHAPTER 3

METHODOLOGY AND DATA COLLECTION

Introduction

In the present study, the researcher has attempted to examine the possible role of language proficiency in critical thinking while reading in the EFL context. This investigation is necessary due to the absence of data concerning how foreign language learners of varying English proficiency levels react to the critical thinking test. So, the following research questions were posed:

1. Is there any meaningful relationship between the levels of language proficiency of EFL students and their performance in English Critical Thinking Test?
2. Is there any meaningful relationship between the levels of language proficiency of EFL students and their performance in Persian Critical Thinking Test?
3. Is there any meaningful difference between the performance of EFL students in English and Persian Critical Thinking Tests?

Just as the groundwork for descriptive study is laid in this chapter and it provides some explanations on the characteristic features of the study, the utmost effort has been devoted to obtain data through the most reliable method. This chapter details the design of the research study undertaken with these goals and questions in mind. The setting of the study including the subjects/participants and materials is described. Then the methodology of the study is detailed including the instruments, general and statistical procedures used. The design section elaborates the process of the data analysis and its meaningful representation. The samples, the variables, the
Validity and Reliability

Validity is that quality of a data-gathering instrument or procedure that enables it to measure what it intends to measure; i.e. a test is valid if it measures what it is supposed to measure. There are several types of validity:

1. Content validity measures the degree to which the test items represent the domain or universe of the trait or property being measured. It refers to the degree to which the test actually measures, or is specifically related to, the traits for which it was designed. The content of the test should be appropriate to what the examinees have learned. It shows how adequately the test samples the universe of knowledge and skills that a student is expected to master. Since it is not possible to have any statistical calculations or numerical method to express content validity; it is usually suggested that a panel of experts in the field to be studied be used to identify a content area. Content validity is particularly important for an achievement test but not very important for aptitude tests.

In the case of the present research, although the TOEFL and Critical Thinking tests were genuine, standard and valid tests, a group of experts in the field of applied linguistics, education and psychology approved the content validity of these tests.

2. Construct or trait refers to what is being measured in a test. Construct validity is the degree to which scores on a test can be accounted for by the explanatory constructs of a sound theory. Construct validity is particularly important for personality and aptitude tests. Computing construct validity involves complicated statistical calculations called Factor Analysis.

Concerning the current study, factorial analysis was run to prove the construct validity of the TOEFL test. The construct validity of the Critical Thinking Test has already been calculated by other researchers according to its manual which has been reported here as well.
3. Criterion-related validity refers to two different types of validity. It is expressed as the coefficient of correlation between test scores and some measure of future performance, or between test scores and scores on another test or measure of known validity.

   a. Predictive validity refers to the usefulness of a test in predicting some future performance.

   b. Concurrent validity refers to the usefulness of a test in closely relating to other measures.

   In this study, the concurrent validity of TOEFL and Critical Thinking tests were measured using the Pearson Coefficient of Correlation.

   Reliability is the degree of consistency that the instrument or procedure demonstrates. The reliability of a research instrument concerns the extent to which the instrument yields the same results on repeated trials. In other words, a test is reliable to the extent that it measures whatever it is measuring consistently, i.e., in case they are repeated, almost the same results are obtained. In tests that have a high coefficient of reliability, errors of measurement have been reduced to a minimum.

   The reliability or stability of a test is usually expressed as a correlation coefficient. There are several types of reliability:

   1. Stability over time (test-retest)

   2. Stability over item samples (equivalent or parallel forms)

   3. Stability of items (internal consistency) which is of two types:

      a. Split halves

      b. Kuder-Richardson formula

   4. Stability over scorers (interscorer reliability)

   5. Standard error of measurement

   Reliability is a necessity but not a sufficient condition for validity. That is, a test must be reliable for it to be valid, but a test can be reliable and still not be valid.

   In the present study, the reliability of all tests was calculated by the use of the Kuder-Richardson formula.
Subjects

The subjects participating in this study were a total of 574 undergraduate Iranian students majoring in English as a foreign language in three fields of English Language Teaching, English Literature and English Translation at Islamic Azad University – Roudehen Branch. Participants were defined as non-native speakers of English who were learning English in a tutored setting. They were males and females at various ages with diverse social classes – since it has been assumed that these variables do not have any impact on the study. Although sex based differences are considered a possible moderating variable and sex differences have been shown to be significant in reading in terms of background knowledge of topics of more interest to one gender than the other (Bügel & Buunk) and in L1 word recognition (Majeres); sex has not been shown to be a significant factor with regard to L2 reading (Brantmeier; Grace; Paivio & Lambert). For this reason it was not considered as a moderating variable in the present study. In addition to the number of students, attempt was made to choose them from different levels of proficiency in order to assign them to three levels of English Language Proficiency, i.e. all of the students were randomly selected from first to eighth semesters. Although participants were selected as intact classes, efforts were made to stress that participation was totally voluntary. Out of an initial pool of 584 students who participated in some phase of the research, 574 completed all the phases of the study. Bases on their performance in the TOEFL test; they were placed at three different proficiency levels: elementary, intermediate and advanced. The subjects were informed of the sensitivity of the research and they cooperated with extensive care, they were rewarded for their efficient participation.

Instrumentation

Considering the purpose of the research, the following tests were developed to measure the role of English language proficiency in Persian and English Critical Thinking when reading. Three instruments of research were manipulated in order to
obtain as valid information as possible. The TOEFL test (Longman, 2005), a standardized and valid test as a measurement for English proficiency was used. A Pilot Test was run in order to validate the test by the researcher himself. This test was administered in order to put the subjects in three proficiency levels of the elementary, intermediate and advanced.

The second instrument was the Cornell Critical Thinking Test, a valid and standard Critical Thinking Test, including 52 test items to elicit and guarantee critical thinking while reading on part of the research subjects.

The third instrument used in the study was the Cornell Critical Thinking Test in Persian. This test is a Persian translation of the Cornell Critical Thinking Test which the experts in the fields of psychology and language learning have already authenticated and the reliability and validity of the test were proven by the researcher.

**Materials**

The materials, exploited during different steps of this research, are all reading passages in the form of MC (multiple choice) items claimed to elicit readers’ critical thinking ability. The proficiency test is a genuine TOEFL, validated and standardized by the researcher. The Cornell Critical Thinking Test (Level Z) is a standard and valid test posed by Ennis in 1985. The Persian version of the Critical Thinking Test was translated and authenticated by experts in the fields of psychology and language learning.

**Cornell Critical Thinking Test**

The Cornell Critical Thinking Test was proposed by Ennis in 1985. This test has been used since 1985 in different fields concerning critical thinking ability. According to its manual, Frisby in 1992 administered the test to 376 Law students and got the reliability of .80. To the same purpose, Solon in 2001 proved the reliability of this test by administering it to 52 students on Humanities and obtaining the reliability
of .74. Ennis and his colleagues administered the test to different groups in different fields in different regions. The results are as follows:

1) 261 students and reliability of .77.
2) 128 undergraduate students in Mid-Western Estate University and reliability of .61.
3) 52 undergraduate students in New England University and reliability of .67.

Solon in his study in 2003 administered the test to 127 students of Linguistics and got the reliability of .72.

Level Z of the Cornell Critical Thinking Test by Spearman-Brown method is as high as .76, .75, .74 and .80. The studies show that the reliability of this test ranges from .49 to .87 for Level Z.

The construct validity of the Cornell Critical Thinking Test is shown in the following table.

<table>
<thead>
<tr>
<th>Section</th>
<th>Number of Items</th>
<th>Construct Validity</th>
</tr>
</thead>
<tbody>
<tr>
<td>I: Deductive Reasoning</td>
<td>10</td>
<td>.76</td>
</tr>
<tr>
<td>II: Semantics and Fallacy</td>
<td>11</td>
<td>.66</td>
</tr>
<tr>
<td>III: Observation and Credibility</td>
<td>4</td>
<td>.60</td>
</tr>
<tr>
<td>IV: Inductive Reasoning</td>
<td>13</td>
<td>.55</td>
</tr>
<tr>
<td>V: Planning Experiments</td>
<td>4</td>
<td>.72</td>
</tr>
<tr>
<td>VI: Definition of Assumption</td>
<td>4</td>
<td>.65</td>
</tr>
<tr>
<td>VII: Identification of Assumption</td>
<td>6</td>
<td>.65</td>
</tr>
<tr>
<td>Total Score</td>
<td>52</td>
<td>.76</td>
</tr>
</tbody>
</table>

Notwithstanding the worldwide application of the Cornell Critical Thinking Test in different studies and educational/academic purposes, this test was not recognized and surveyed in Iran. So, the researcher endeavored to investigate the reliability and validity of the test in Iranian students majoring in English Language.

On the other hand, the correlation coefficients of this test obtained by different studies are demonstrated in the following table.
Moreover, the Persian version of the Critical Thinking Test was translated and authenticated by the experts in the fields of psychology and language learning in Iran. Alivandi in Iran administered the Persian version of the Cornell Critical Thinking Test to 249 students of different fields in Tabriz University in order to investigate the relationship between Critical Thinking and Academic Achievement. She obtained the Pearson Correlation Coefficient of .128 which was significant in her study, i.e. there is a significant relationship between critical thinking and academic achievement. Furthermore, by doing test-retest experiment, she gained the reliability of .71 for Persian version of the Cornell Critical Thinking Test. Supplementary; the researcher obtained the KR-21 reliability of .88 and Pearson Correlation Coefficient of .85 for Persian Cornell Critical Thinking Test.

**Proficiency Levels for Reading/Critical Thinking**

**Level 1**

At level 1, a student can:

- Recognize factual material explicitly presented in a reading passage.

<table>
<thead>
<tr>
<th>Correlation</th>
<th>Number of Subjects</th>
<th>Group / Field</th>
<th>Researcher and Year</th>
<th>Correlation with</th>
</tr>
</thead>
<tbody>
<tr>
<td>.71</td>
<td>100</td>
<td>Undergraduate Students of Social Science and Mathematics</td>
<td>Minse 1990</td>
<td>Watson-Glaser Critical Thinking Test</td>
</tr>
<tr>
<td>.53</td>
<td>25</td>
<td>Undergraduate Students</td>
<td>Royalty 1995</td>
<td>Statistical Reasoning</td>
</tr>
<tr>
<td>.67</td>
<td>92</td>
<td>University Teachers</td>
<td>Phelps 1987</td>
<td>Nelson Mental Ability</td>
</tr>
<tr>
<td>.51</td>
<td>20</td>
<td>Freshmen</td>
<td>Hill 1995</td>
<td>Moral Judgment Interview</td>
</tr>
<tr>
<td>.62</td>
<td>20</td>
<td>Freshmen</td>
<td>Hill 1995</td>
<td>Educational Achievement</td>
</tr>
<tr>
<td>.30</td>
<td>165</td>
<td>Undergraduate Students (Freshmen)</td>
<td>Farely 1984</td>
<td>Cognitive Development Ability Test</td>
</tr>
<tr>
<td>.46</td>
<td>100</td>
<td>Graduate and Undergraduate Students</td>
<td>Minse 1990</td>
<td>Logical Judgment Model</td>
</tr>
</tbody>
</table>
Level 2
At level 2, a student can:
- Synthesize material from different sections of a passage.
- Recognize valid inferences derived from material in the passage.
- Identify accurate summaries of a passage or of significant sections of the passage.
- Understand and interpret figurative language.
- Discern the main idea, purpose or focus of a passage or a significant portion of the passage.

Level 3
At level 3, a student can:
- Evaluate competing casual explanations.
- Evaluate hypotheses for consistency with known facts.
- Determine the relevance of information for evaluating an argument or conclusion.
- Determine whether an artistic interpretation is supported by evidence contained in a work.
- Recognize the salient features or themes in a work of art.
- Evaluate the appropriateness of procedures for investigating a question of causation.
- Evaluate data for consistency with known facts, hypotheses or methods.
- Recognize flaws and inconsistencies in an argument.

Data Collection and Analysis Procedure

The procedure of the present research to collect data has been conducted in the following phases:
First Phase

In order to validate and standardize the TOEFL test (Longman, 2005), a pilot study was run. A TOEFL test with 90 items was administered to undergraduate students majoring in English at Islamic Azad University – Tabriz Branch. 77 students were invited to attend the research by their instructors and the researcher. The following statistical analyses are employed in the pilot study to analyze the obtained data.

A- Item Analysis

Item facility (IF) and item discrimination (ID) and the KR-21 reliability coefficient are the indices calculated to evaluate the items of the instruments employed in this study. This section includes a brief introduction to these concepts, their computation and interpretation.

1- Item Facility (IF)

Item facility refers to the degree of the easiness of an item. It is simply calculated by dividing the number of cases who have answered the item correctly by the total number of subjects.

\[
IF = \frac{\sum C}{N}
\]

IF: Item Facility  
\(\sum C\): Total number of correct answers  
N: Total number of subjects

Item facility ranges from 0 to 1. An IF of zero indicates that all of the subjects have missed the item and the item is absolutely difficult. On the other hand, an IF of 1 shows that all of the subjects have answered the item correctly. In other words, the item is rather easy.

An ideal item facility should be equal or higher than .37 and equal or lower than .63.

\[.37 \leq IF \leq .63\]
2 - Item Discrimination (ID)

A good item should be able to differentiate between the weak and strong students. In other words, an item enjoys discriminatory power if it is answered by the strong students while the weak students miss it. In order to calculate the Item discrimination (ID) the following steps should be taken:

a. Compute the total score for each subject.
b. Sort the subjects on their total scores.
c. Choose the upper and lower 25 percent students (In some references 33 and 37 percents are also mentioned).
d. Tally the correct answers in both groups and plug them into the following formula:

\[ ID = \frac{\sum CH - \sum CL}{N / 2} \]

Item discrimination can range from -1 to 1. An ID of -1 shows that all the weak students have answered the item correctly, while all the strong students have missed it. An ID of 1 shows the reverse case. That is, all the strong and none of the weak students have answered the item correctly. An ID of zero indicates that the item cannot differentiate strong and weak students. In other words, the strong and weak students may choose the correct answer.

An ideal ID should be equal or higher than .40.

\[ ID \geq .40 \]
3- Kuder-Richardson's Reliability Coefficient (KR-21)

The concept of reliability refers to the degree of the consistency of the test scores. In other words, a good test should yield the same results if it is replicated. A number of different formulas have been devised to compute the reliability coefficient. The KR-21 is one of the most widely used and easily calculated indices. All one needs are the mean, variance and the number of the items in the test to calculate the index through the following formula:

$$KR-21 = \left[ \frac{K}{K-1} \right] \left( 1 - \frac{\bar{X} - (K - \bar{X})}{KV} \right)$$

KR-21 Kuder-Richardson-21 Reliability Coefficient
K Number of items in the test
\(\bar{X}\) Mean score
V Variance of the test scores

Reliability coefficient may range from 0 to 1, the former shows that the test is not reliable at all, while the latter shows perfect reliability of the test. For a teacher-made test, a reliability coefficient of .60 or more is good, while for a standardized test it should be .80 or more.

Second Phase

Seventeen classes were randomly selected as a sample from the different levels of education, i.e. undergraduate Iranian students majoring in English Language Teaching, English Literature and English Translation at Islamic Azad University - Roudehen Branch. 574 students were invited to attend the research by their instructors and the researcher. A standardized and valid TOEFL test was administered as a proficiency test (App.: I) to the subjects of research in order to form three groups of subjects at three levels of proficiency, i.e., elementary, intermediate and advanced.
Since, the reading process is involved in this research project and the researcher endeavors to investigate the role of language proficiency in critical thinking when reading, TOEFL contained two sections with 90 items:

I. Structure and Written Expressions (40 multiple-choice items)
II. Reading Comprehension (50 multiple-choice items)

The answer sheets gathered were rated objectively according to the answer key provided. Each student’s raw score was revised to its converted and standard score according to the Conversion Chart available in Longman TOEFL (App.: II). Hence, students were placed in three different levels of proficiency on the basis of their scores obtained from TOEFL so that those with scores obtained between + 0.5 SD above and – 0.5 SD below the mean score were considered as the intermediate group. Those subjects who obtained beyond + 0.5 SD above the mean score and – 0.5 SD below the mean score were considered as the advanced and elementary subjects respectively. One-way Analysis of Variances (One-way ANOVA) was run to compare the mean scores of the three groups and indicate their significant differences.

One-way Analysis of Variances (One-way ANOVA)

The statistical analysis of variances are developed in order to overcome the shortcomings of the t-test. Although the t-test is robust test, it cannot be used to compare the mean scores of all of the groups participating in a study. The maximum number of comparisons that can be made using the t-test is one minus the number of groups under the study. Moreover the t-test cannot be used to cross-compare the groups. For example if there are four groups in a study, by using the t-test, one can make only three comparisons, provided that none of the comparisons is crossed. That is to say, after comparing the mean score of group 1 and 2, and group 2 with group 3, one can not compare group 1 with group 3.

The family of statistical techniques known as the analysis of variances has been developed to cope with situations in which more than two groups are involved and one needs to compare all of the groups simultaneously.

One-way analysis of variance is one such statistical test that can be used to compare the mean scores of more than two groups participating in one test. Thus there
are two variables involved in each one-way ANOVA; a dependent variable which consists of the test scores and an independent variable which is used to code the groups in the study. There are three proficiency groups in this study; advanced, intermediate and elementary. The research question involves comparing the mean scores of the three groups on the critical thinking test. The computation can be carried out through the following steps:

1- **Sum of Squares Total (SST)**

The sum of the squared scores should be computed first:

$$SST = \sum X^2 - \frac{(\sum X)^2}{N}$$

- **SST**: Sum of Squares Total
- $$\sum X^2$$: Square the scores one by one and then sum them up
- $$(\sum X)^2$$: Sum the scores up. Then square this number
- **N**: Total number of subjects

2- **Sum of Squares Between (SSB)**

The right-hand section of the above formula should be computed for each group separately at the second stage:

$$SSB = \left[ \frac{(\sum X_1)^2}{n_1} + \frac{(\sum X_2)^2}{n_2} + \ldots + \frac{(\sum X_k)^2}{n_k} \right] - \frac{(\sum X)^2}{N}$$

- **SSB**: Sum of Squares Between
- $$(\sum X_1)^2, (\sum X_2)^2, \ldots$$: Sum the scores of the first group and then square it. Do the same for other groups
Number of subjects in the 1st, 2nd, 3rd, etc. groups

\( \sum x \) - Sum the scores of all groups up. Then square this number

N - Total number of subjects

3- **Sum of Squares Within (SSW)**

Sum of squares within is the difference between the SST and the SSB.

\[
SSW = SST - SSB
\]

4- **Mean Square Between (MSB)**

Mean square between can be computed through the following formula the elements of which are explained in the above sections.

\[
MSB = \frac{SSB}{K - 1}
\]

5- **Mean Square Within (MSW)**

Mean square within can be computed through the following formula the elements of which are explained in the above sections.

\[
MSW = \frac{SSW}{N - K}
\]

6- **F-value**

All remained is to compute the F-value as follows:

\[
F = \frac{MSB}{MSW}
\]
7- Interpretation

After calculating the F value, one needs to determine its statistical significance in one of the following two methods:

7-1- The easiest method is to check the significance level that is produced by the Statistical Package for the Social Sciences (SPSS) together with the F-value. If the significance level is equal or lower than .05, it can be concluded that there are significant differences among the means of the different groups. On the contrary, if the significance level is higher than .05 one can claim that the differences among the means are not statistically meaningful.

7-2- The second method involves comparing the observed F-value with the critical F-value which can be found under statistical tables. In order to find the critical F-value, one needs two sets of degrees of freedom (D.F.): between groups and within groups. The between groups degree of freedom equals the number of groups minus 1. For example, if comparison is made among the mean scores of three groups on a test, then the between groups degree of freedom will be 2 (3-1=2).

The second degree of freedom is the within groups D.F which is calculated by subtracting the number of groups from the total number of subjects. For example if 574 subjects from 3 different groups participated in an exam, the within groups degree of freedom will be 571 (574-3=571).

Having calculated these two sets of degrees of freedom, one can refer to the statistical tables and find the critical F-value which corresponds to these two degrees of freedoms. If the F-observed value is equal or higher than the critical F-value, one can conclude that there are significant differences among the means. On the other hand, if the observed F-value is lower than the critical value of F, then one should come to the conclusion that there is not any significant difference among the means.

Post Hoc Comparisons

A significant F-observed value indicates that there are significant differences among the means; however, it does not show which pairs of means are significantly different and which are not. In order to compare any pairs of means, one can use post hoc comparison tests. Two such tests that are widely used are Scheffe's Test and
Tukey Test. The former is more conservative and declares large differences as significant, while the latter is less conservative and may declare marginal differences as significant.

In this study, the Scheffe’s post hoc comparison tests were employed to compare the pairs of means after running the one-way ANOVA.

The post hoc Scheffe’s test may simply be calculated through the following formula:

$$|\bar{X}_1 - \bar{X}_2| \geq MSW \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- $|\bar{X}_1 - \bar{X}_2|$ Absolute value of the difference between the two means
- $MSW$ Mean Square Within
- $n_1, n_2$ Number of subjects in the 1\textsuperscript{st} and 2\textsuperscript{nd} groups

**Third Phase**

As far as critical thinking skills and strategies were concerned for the purpose of the study, the Cornell Critical Thinking Test (Level Z) was adopted (App.: III). The Cornell Critical Thinking Tests develop a clear picture of students’ critical thinking abilities. The tests can be used to test critical thinking skills, to predict students’ performance in proficiency exam, critical thinking courses, college admissions, career and employment, methods and teaching support. Level X includes induction, deduction, credibility and identification of assumptions. Level Z contains the skills in Level X plus semantics, definition and prediction in planning experiments.

This test contains 7 sections and 52 multiple-choice items. The first section covers deductive reasoning and contains 10 items. In the second section Ennis has adopted semantics and fallacy containing 11 items. Observation and credibility are included in the third section with 4 items. The fourth section takes advantage of inductive reasoning with 13 items. In the fifth section students’ abilities in planning
experiments are tested through 4 items. The sixth and seventh sections include definition of assumption and identification of assumption with 4 and 6 items respectively.

The Cornell Critical Thinking Test (Level Z) was administered to 574 students in three levels of language proficiency, i.e. elementary, intermediate and advanced. Again the answer sheets obtained were rated objectively according to the answer key provided. The Pearson Correlation Coefficient was employed to calculate the validity index of the English Critical Thinking Test.

**Pearson Coefficient of Correlation**

This statistic is used to measure the degree of relationship between two variables. The Pearson ‘R’ measures the relationship between two sets of scores. If one of the sets is an already validated variable, then the Pearson ‘R’ shows the validity of the other set. If it can be proved that the English Critical Thinking Tests have high correlation with the TOEFL whose validity has been approved, then it can be claimed that the critical thinking test is valid as well.

1. **Computation**

The Pearson Correlation Coefficient is calculated with the following formula:

\[ r = \frac{N(\sum XY) - (\sum X)(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}} \]

- \( r \) Pearson Correlation Coefficient
- \( X \) Subjects Scores on the First Test
- \( Y \) Subjects Scores on the Second Test
- \( N \) Number of subjects
2. Interpretation

The interpretation is almost the same as the one presented for One-way ANOVA. Either one has to check the significance of the r-value, i.e. if the significance is lower than .05, one can claim that there is a meaningful relationship between the two tests. The other method is to check the observed t-value with the critical value of the Pearson Coefficient.

After approving the reliability and validity of English Critical Thinking Test, a one-way ANOVA was run to compare the mean scores of the three proficiency groups on English Critical Thinking Test to conclude whether there is a meaningful relationship between the levels of language proficiency of EFL students and their performance in English Critical Thinking Test.

At the end of this phase the post hoc Scheffe’s test was adopted to compare any pairs of means, i.e. to conclude whether there are significant differences between any levels of students’ language proficiency and their performance in English Critical Thinking Test.

Fourth Phase

Since the purpose of the present study is to investigate the role of language proficiency in critical thinking considering reading processes in EFL context, the researcher scrutinized a comparative study of critical thinking test both in English and Persian while reading. Accordingly, at this phase, the Persian version of the Cornell Critical Thinking Test (App.: IV) was administered to the same 574 students in three levels of language proficiency. The answer sheets were rated objectively according to the answer key provided. The Pearson Correlation Coefficient was employed to calculate the validity index of the Persian Critical Thinking Test.

After approving the reliability and validity of the Persian Critical Thinking Test, a one-way ANOVA was adapted to compare the mean scores of the three proficiency groups on the Persian Critical Thinking Test to establish if there is a meaningful relationship between the levels of language proficiency of EFL students and their performance in Persian Critical Thinking Test.
In the last part of this phase the post hoc Scheffe’s test was applied to compare any pairs of means, i.e. to probe that there are significant differences between any levels of students’ language proficiency and their performance in Persian Critical Thinking Test.

**Fifth Phase**

In the last phase of the current study, the mean scores of the students in the English and Persian Critical Thinking Tests were compared in order to explore whether the students are better critical thinkers in English or in Persian while reading. Consequently, Paired Samples t-test was adopted to establish the differences.

**Paired Samples t-test**

The paired samples t-test statistical analysis is run to compare the mean scores of a single group on two different tests. In case of this study, paired samples t-test is used to compare the mean scores of the English and Persian Critical Thinking Tests.

**1. Computation**

The formula for calculating the t-statistic is as follows:

$$t = \frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{\frac{\sum D^2 - (1/N)(\sum D)^2}{N - 1}} \sqrt{N}}$$

- $|\bar{X}_1 - \bar{X}_2|$: Absolute value of the difference between the two means
- $\sum D^2$: Sum of the squared differences between the two scores
- $\sum (D)^2$: Sum up the differences between the two tests. Then square it.
- $N$: Number of subjects
2. Interpretation

The interpretation of the paired samples t-test results is the same as for the analysis of variance. The t-observed value can be interpreted in either of the following ways:

2-1- The easiest method is to check the significance level that is produced by the SPSS together with the t-value. If the significance level is equal or lower than .05, it can be concluded that there is a significant difference between the mean scores of the two tests. On the contrary, if the significance level is higher than .05 one can claim that the difference between the means is not statistically meaningful.

2-2- The second method involves comparing the observed t-value with the critical t-value which can be found under statistical tables. In order to find the critical t-value, one needs the degree of freedom (D.F.) which equals the number of the subjects minus 1. For example, if 574 subjects have performed on the two tests, then the degree of freedom is 573.

The Design of the Study

Since the nature of the present study is a correlational analysis, the design of the current research is ex post facto research.

The application of an ex post facto design is vindicated on the grounds that the distinction between the dependent and independent variables appears to be arbitrary rather than rule-governed. Since the events or conditions have already occurred or existed, the researcher merely selects the relevant variables for an analysis of their relationship. Any relationship between the scores of the groups would not be related to any instructional program. The researcher takes the effect (or dependent variable) and examines the data retrospectively to establish causes, relationships or associations and their meanings. The researcher has no control over what has already happened to the subjects. The treatment, whatever it might be, is given prior to the research project. There is no cause-and-effect concern between variables, only the type and/or degree of their relationship is under question. Ex post facto design just seeks to find answers to questions through an analysis of variable relationships. Hatch and Farhady
support this by stating that "ex post facto design often used when the researcher does not have control over the selection and manipulation of the independent variable." On the other hand, Best and Kahn prompt the researcher that s/he must be attentive to the fact that the information used in ex post facto studies may be incomplete. The researcher may not have sufficient information about all of the events or variables occurring at the time of study. Other features of ex post facto research become perceptible when it is contrasted with true experimental research. Basically, ex post facto investigations have a built-in weakness: lack of control of the independent variable or variables. According to Spector, it is impossible to isolate and control every possible variable, or to know with absolute certainty which are the most crucial variables.

Cohen, Manion and Morrison believe that it should not be concluded that ex post facto studies are of little value; many important investigations in education and psychology are ex post facto designs. Most of the time there is no choice in the matter; a researcher cannot cause one group to become failures, delinquent, etc. Research necessarily should rely on existing groups. Alternatively, the inability of ex post facto design to incorporate the basic need for control makes them vulnerable from a scientific point of view and the possibility of their being misleading should be clearly acknowledged. Ex post facto designs are probably better realized not as experiments with the greater certainty but more as surveys, useful as sources of hypotheses to be tested by more conventional experimental means of a later date. This research is particularly suitable in social, educational and psychological contexts where the independent variable or variables lie outside the researcher’s control.

Consequently, one may conclude that ex post facto designs are appropriate in circumstances where the more powerful experimental method is not possible. As Cohen, Manion and Morrison (207-8) declare “These would arise when it is not possible to select, control and manipulate the factors necessary to study cause-and-effect relationships directly; or when the control of all variables except a single independent variable may be unrealistic and artificial, preventing the normal interaction with other influential variables; or when laboratory controls for many research purposes would be impractical, costly or ethically undesirable.”

Spector (43) identifies the advantages of ex post facto research as follows:
Ex post facto research meets an important need of the researcher where the more rigorous experimental approach is not possible.

The method yields useful information concerning the nature of phenomena – what goes with what and under what conditions. In this way, ex post facto research is a valuable exploratory tool.

Improvements in statistical techniques and general methodology have made ex post facto designs more defensible.

In some ways and in certain situations the method is more useful than the experimental method, especially where the setting up of the latter would introduce a note of artificiality into research proceedings.

Ex post facto research is particularly appropriate when simple cause-and-effect relationships are being explored.

The method can give a sense of direction and provide a fruitful source of hypotheses that can subsequently be tested by the more rigorous experimental method.

Among the limitations and weaknesses of ex post facto designs Spector (43) explains the following:

- There is the problem of lack of control in that the researcher is unable to manipulate the independent variable or to randomize her subjects.
- One cannot know for certain whether the causative factor has been included or even identified.
- It may be that no single factor is the cause.
- A particular outcome may result from different causes on different occasions.
- When a relationship has been discovered, there is the problem of deciding which is the cause and which the effect; the possibility of reverse causation has to be considered.
- The relationship of two factors does not establish cause and effect.
- Classifying into dichotomous groups can be problematic.
- There is the difficulty of interpretation and the danger of the ex post facto assumption being made, that is, believing that because X precedes O, X causes O.
- It often bases conclusions on too limited a sample or number of occurrences.
• It frequently fails to single out the really significant factor or factors, and fails to recognize that events have multiple rather than single causes.
• As a method it is regarded by some as too flexible.
• It lacks nullifiability and confirmation.
• The sample size might shrink massively with multiple matchings.

As one of the most commonly used subsets of ex post facto design, the researcher used a correlational analysis. This inter-relational descriptive research method enables the researcher to find the degree of relationship between the independent variable that is language proficiency in this study and the dependent variable, which is critical thinking. The independent variable consists of three levels, defined as elementary, intermediate and advanced. Participants were assigned to one of the three levels of language proficiency based on their scores on the TOEFL test. The design was a 2 by 3 factorial design; that is the presence or absence of critical thinking in three levels of proficiency.