Chapter 2

Methodology

This chapter presents the methods used in the study. These methods included the information on study design, population and samples, variables, data collection, management and data analysis, and methods for statistical analysis.

2.1 Study Design

The cross-sectional study was performed in the second semester of the academic year 2005 among the undergraduate students at Prince of Songkla University, Pattani Campus.

2.2 Population and Samples

The population was the undergraduate students studying in the academic year 2005 at Prince of Songkla University, Pattani Campus.

The samples comprised 785 undergraduate students studying in the second semester of the academic year 2005 at Prince of Songkla University, Pattani Campus. They were selected by purposive sampling, classified by their major subjects; (1) sciences and (2) arts, from the faculties of Education, Humanities and Social Sciences, Science and Technology, Communication Science, Fine and Applied Arts, and the College of Islamic Studies as shown in Table 2.1.

Major subjects	Samples
Sciences	344
Arts	441
Total	785

Table 2.1: Samples selected

2.3 Variables

Determinants

The determinants in the study were gender, religion, degree duration, seniority level, faculty and major field of study, high school type and home province. In addition, the time taken to complete the English Vocabulary Skill Test was also consisted as a determinant.

Outcome

The outcome was the English Vocabulary Skill. This outcome was measured by the English Vocabulary Skill Test scores, ranged from 0 to 10. The English Vocabulary Skill Test is appeared in the Appendix A.

The schematic diagram for this study is shown in Figure 2.1.

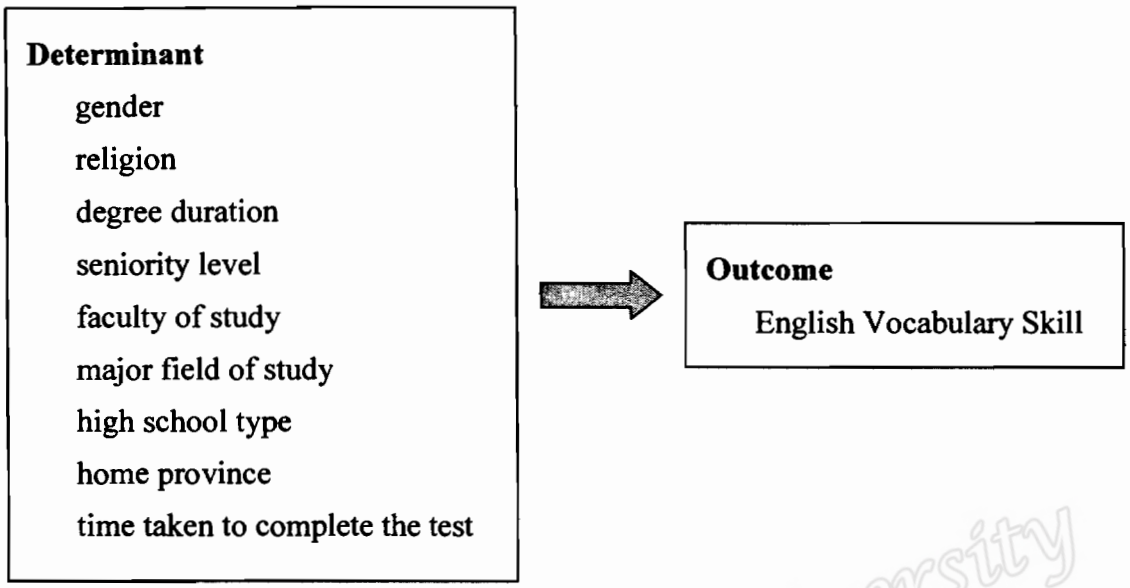

Figure 2.1: Schematic diagram of variables of interest

2.4 Data collection, management and data analysis

Data collection

The instruments used in collecting data were an English Vocabulary Skill Test and the university MIS (Management Information System) database.

The English Vocabulary Skill Test was constructed to measure the English vocabulary skill of the undergraduate students. The test contained 10 common English words gathered from the widely-used English textbooks at high schools in Thailand and introductory Statistics text at Macquarie University in Australia. For each item, there were 5 possible response choices: a correct synonym, a similar sounding word, a similarly written word, an opposite word, and another unrelated word listed in alphabetical order next to the test item. These items were listed in random order. To complete the test, the students were requested to select the word that most closely matched the given word in meaning. Then, the answer was specified as 1 score if

correct and 0 score if incorrect. Furthermore, the students were additionally invite without coercion to give their students' ID number, enabling demographic and enrolment detail to be matched to their answers and also the time they started and finished the test. The students' ID number subsequently was used to link via the university MIS (Management Information System) database to identify the gender, religion, degree duration, seniority level, faculty and major field of study, high school type, and home province.

Therefore, the English Vocabulary Skill Test was distributed to a sample of the undergraduate students studying in the academic year 2005 from the faculties of Education, Humanities and Social Sciences, Science and Technology, Communication Science, Fine and Applied Arts, and the College of Islamic Studies at Prince of Songkla University, Pattani Campus.

Data management

The data were recorded in a Microsoft Excel spreadsheet file, imported to Microsoft SQL Server and then analyzed using *Webstat* (a set of programs for graphical and statistical analysis of data stored in an SQL database, written in HTML, VBScript).

Data analysis

Analysis of variance and two sample t-tests were used for preliminary analysis. Since the outcome variable was continuous and the determinants comprised more than one variable, multiple regression analysis was the appropriate method for statistical modeling.

2.5 Statistical methods

This study focused on the association between the English Vocabulary Skill and the gender, religion, degree duration, seniority level, faculty of study, major field of study high school type, home province and time taken to complete the English Vocabulary Skill Test of the undergraduate students at Prince of Songkla University, Pattani Campus.

Two sample t-test

The two sample t-test is used to test the null hypothesis that the two population means are the same. The t-statistics is obtained as follows

$$t = \frac{\bar{y}_1 + \bar{y}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{2.1}$$

If s_1 and s_2 denote the standard deviations of the two samples, respectively, it may be shown that the pooled sampled standard deviation is given by the formula

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{2.2}$$

A p-value is now obtained from the table of the two-tailed t distribution with $n_1 + n_2 - 2$ degree of freedom (McNeil, 2005).

One-way analysis of variance

Considering the analysis of data in which the outcome is continuous and the determinant is categorical, this leads to a procedure called the (one-way) analysis of variance (ANOVA). The null hypothesis is that the population means of the outcome

variable corresponding to the different categories of the determinants are the same, and this hypothesis is tested by computing a statistic called the *F-statistics* and comparing it with an appropriate distribution to get a *p*-value. Suppose that there are n_j observations in sample j denoted by y_{ij} for $i = 1, 2, \ldots, n_j$. The *F*-statistics is defined as

$$F = \frac{(S_0 - S_1)/(c-1)}{S_1/(n-c)} \tag{2.3}$$

where $$S_0 = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(y_{ij} - \bar{y})^2, S_1 = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2$$

and $$\bar{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}, \bar{y} = \frac{1}{n}\sum_{j=1}^{c}\sum_{i=1}^{n_j} y_{ij}, n = \sum_{j=1}^{c} n_j$$

S_0 is the sum of squares of the data after subtracting their overall mean, while S_1 is the sum of squares of the residuals obtained by subtracting each sample mean. If the population means are the same the numerator and the denominator in the *F*-statistics are independent estimates of the square of the population standard deviation (assumed the same for each population) and the *p*-value is the area in the tail of the *F*-distribution with *c*-1 and *n-c* degrees of freedom (McNeil, 1996).

Multiple linear regression

Regression is used to analyse data in which the outcome is continuous variable. If there is a single determinant, the data may be displayed as a scatter plot and summarized by fitting a straight line. In conventional statistical analysis the line fitted is the *least squares line*, which minimizes the distances of the points to the line, measured in the vertical direction. If there is more than one determinant, the method

generalises to multiple linear regression, in which the regression line extends to the multiple linear regression represented as

$$Y = \beta_0 + \sum \beta_i x_i + \varepsilon \tag{2.4}$$

where Y is the outcome variable, β_0 is a constant, $\{\beta_i\}$ is a set of parameters (i = 1 to p), and $\{x_i\}$ is a set of determinants (i = 1 to p) (McNeil, 1996).

The model is fitted to data using least squares, which minimises the sum squares of the residuals.

Linear regression analysis resets on three assumptions as follows.

(1) The association is linear.

(2) The variability of the error (in the outcome variable) is uniform.

(3) These errors are normally distributed.

If these assumptions are not met, a transformation of the data may be appropriate. Linear regression analysis may also be used when one or more of the determinant is categorical. In this case the categorical determinant is broken down into c-1 separate binary determinants, where c is the number of categories. The omitted category is taken as the baseline or referent category.

Correlation Coefficient

The correlation coefficient is a measure of the linear or straight-line, relationship between variables and level of relation. The correlation coefficient is defined as (McNeil, 2005)

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \tag{2.5}$$

Residual analysis

In conventional linear regression models where errors are assumed to be independent and normally distributed, the adequacy of the model can be assessed by plotting residuals, obtained from the observations simply by subtracting their conditional means, against corresponding normal scores. If the data have a normal distribution, the normal score plot should show a linear trend. It is assumed that the normality assumption is appeared to be reasonable.