Chapter 2

Methodology

In this chapter we describe the methods for selecting the subjects from the target population and for collecting the data from the subjects and statistical methods used to analyse the data.

2.1 Study Design and Sampling Technique

A cross-sectional study design is used for this study. The target population comprises high school students in Songkhla province. Sampled groups were students in high school classes Ms 4, Ms 5, and Ms 6 in year 2000 at 10 high schools in Songkhla province. The sample size required for specified precision d is given by the formula (McNeil, 1996: 267)

$$n = \frac{z_{\alpha/2}^2 \sigma^2}{d^2},$$

where σ is the standard deviation of the measurement and $z_{\alpha/2}$ is the critical value of the normal distribution separating a symmetric tail area α.. For a 95 % confidence interval, $\alpha = 0.05$ so

$$z_{\alpha/2} = 1.96.$$

Srisuyo (1984) studied determinants of learning style in teaching colleges in Bangkok. She estimated the variance of responses to be 0.1119, and if we assume the required precision is 0.05, the formula gives

$$n = \frac{1.96^2(0.1119)}{0.05^2} \approx 172 \text{ students}$$

So a sample size of 172 is needed. However the sample size is this study 196 are obtained. The sample should be representative the population. In this study, the stratified random sampling method is used to get a sample. The samples are grouped in each program by school size, as shown in Table 2.1.

program School name	Science/Maths			Engl/Maths			Engl/France			General			Total
	Ms4	Ms5	Ms6	Ms4	Ms5	Ms6	Ms4	Ms5	Ms6	Ms4	Ms5	Ms6	
Large													
Hatyaiwitayalai	4	4	4	-	-	-	5	5	5	-	-	-	27
Hatyairatprachasan	-	-	-	2	2	2	-	-	-	2	2	2	12
Woranarichalerm	-	-	-	-	-	-	5	5	5	2	2	1	20
Mahawachirawut	4	4	4	2	2	2	-	-	-	-	-	-	18
Medium-Large													
Songkhlawittayakom	-	-	-	-	-	-	-	-	-	4	3	3	10
Rattapum wittaya	4	4	3	3	3	2	-	-	-	-	-	-	19
Hatyaiwitayalai 2	4	3	3	3	3	2	-	-	-	-	-	-	18
Medium													
Sabayoiwittaya	4	3	3	-	-	-	-	-	-	6	7	6	29
Ranotwittaya	4	4	4	6	6	6	-	-	-	-	-	-	30
Small													
Rattanapolwittaya	5	0	0	-	-	-	-	-	-	4	4	-	13
Total	72			46			30			48			**196**

Table 2.1: The sample size for each school

The variables of interested for this study comprise 9 determinants and 6 outcomes. The details of the determinant variables are as follows:

Variables	Definition
gender	Male and female
grade	Class year
program	Science-Mathematics, English-Mathematics, English-French, General
GPA	the grade-point average score of each student in 1st class of year 1999, divided into 3 levels
SchoolSize	Size of school
Foccup	Father'Occupation
MOccup	Mother'Occupation
FamInc	Father and mother's income per month
Nsibs	Number of brothers and sisters

Outcome Variable

The outcome variables have six components, as follows:

(a) *Independence*: the students like to think by themselves, usc their imagination, and strongly believe in their sclf-confidence.

(b) *Avoidance*: The style that thc students will pay no attention in subject, no participation in learning activity, and not interested in what going on in classroom.

(c) *Collaboration*: The students like to work with other students, try to join and cooperate with other peoplc in learning activity. They believe that classroom is thc proper place to contact society and to learn subjects.

(d) *Dependence*: The students like to study only the subjects that arc assigned to them. Teachers and friendly classmatcs arc the source of knowledge and can help them. They arc ready to receive orders telling what they have to do.

(e) *Competition*: In this style the students try to do everything to be bctter than the others in classroom. They consider the classroom as a competitive atmosphere and they have to win all the time.

(f) *Participation*: Thc students like to go to class. They realize that thcy have to participate in class in every activity, Howcver, they will even participate in activities outside the normal scope.

2.2 Graphical and Statistical Methods

The graphical methods are presented and obtained by using Matlab program version 5 (Hanselman and Littlefeld, 1997) and EcStat 2000, as follows:

1. Histograms and numerical summaries for data from all variables.
2. Factor analysis, two-sample t-test, one way, two way analysis of variance and Multiple regression analysis of the variables described by box plots and 95% confidence intervals of means.

Factor Analysis

Since we have multivariate outcomes, factor analysis is used to reduce the dimensionality of these outcomes. Factor analysis is a data reduction technique. It is a group of procedures designed for removing duplicated information from a set of correlated variables and representing the variables with a smaller set of derived variables or factors. There are three procedures involved. The first stage is obtaining the original data matrix. A set of subjects 0_1, 0_2,....,0_n are measured with a different number of variables V_1,V_2,....., V_k. The second stage involves the creation of a correlation matrix, which is calculated for each combination of two variables: V_1 with V_2, V_1 with V_3, etc, according to the following formula:

If x_i is the data from V_1, and y_i is the data from V_2, then the correlation is given by

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

where s_x and s_y are the sample standard deviations of V_1 and V_2, and n is the number of pairs of observations.

The last stage involves computing the factor loadings. These reveal the extent to which each of the variables contributes to the meaning of each factor. Within any one column of the factor matrix, some of the loadings will be high and some will be low. The variables with the high loading on a factor will be the ones that provide the meaning of the factor (Kachigan, 1991).

Maximum likelihood factor analysis is a widely used method. This method enables us to carry out test of the goodness of fit of a solution comprising *k* factors. It provides a test of the null hypothesis that *k* common factors are sufficient to describe the data. The algorithms for this method are given as follows.

Suppose we have *p* variables and want to fit *k* factors. Let R be the $p \times p$ correlation matrix of the variables, *L* the $p \times k$ matrix of factor loadings, and ψ the vector of length *p* containing the unique variances. Then we need to find values for *L* and ψ that maximise the likelihood function, $f(L,\psi)$.

For the fixed value of ψ, we maximize $f(L,\psi)$ with respect to *L*. The value of *L* is then substituted into $f(L,\psi)$. Now f can be reviewed as a function of ψ. A transformation of this function gives

$$m(\psi) = \sum_{m=k+1}^{p} \left[\log \gamma_m + \frac{1}{\gamma_m} - 1 \right]$$

where $\gamma_1 \le \gamma_2 .... \le \gamma_p$ are the eigenvalues of $\psi R^{-1} \psi$. We then minimize $m(\psi)$. This gives an estimate of ψ, which is then put into the likelihood $f(L,\psi)$. Then the likelihood is again maximized with respect to *L*. Then a new value for $m(\psi)$ is computed and so on.

After making the decision on how many factors to extract from the original set of variables we can redefine the factors so that the explained variance is redistributed among the new factors. This technique is used to sharpen the distinction in the meaning of the factors. A redefinition of the factors, with the loading on the various factors either very high or very low, and then eliminating as many medium sized loading, aids in the interpretation of factors.

Varimax rotation is one of many types of rotation and is regarded as the standard approach. This approach places more emphasis on the simplification of the factors. It tends to avoid a general factor. Using the comprehensibility method to select a number of factors, suppose that three factors are retained. Table 2.2 shows the factor loadings before and after using a rotation of the factors.

Before rotation			
Variable	F_1	F_2	F_3
V_1	M	M	L
V_2	H	L	L
V_3	M	M	L
V_4	M	L	H
V_5	H	M	M
V_6	H	M	M
V_7	L	H	M
V_8	M	M	H
V_9	M	M	L

After rotation			
Variable	F_1	F_2	F_3
V_1	M	M	H
V_2	H	L	L
V_3	L	H	L
V_4	L	L	H
V_5	H	L	M
V_6	H	M	L
V_7	M	H	L
V_8	L	M	H
V_9	H	L	L

Factor loading H : high, M : median and L : low

Table 2.2: Factor rotation

Two sample t-test

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

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

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

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

A p-value is now obtainable from the table of the two-tailed *t* distribution with $n_1 + n_2 - 2$ degrees of freedom. This statistical procedure is called the *two-sample t test* (McNeil, 2000: p 171-172).

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 determinant are the same, and this hypothesis is tested by computing a statistic called the *F-Statistic* 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, ..., n_j. The F-statistic is

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

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

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-statistic are independent estimates of the square of the population standard deviation (assumed the same for each population). The p-value is the area in the tail of the F-distribution with *c*–1 and *n*–*c* degrees of freedom (McNeil, 1996: pp 67, 73).

Two-way analysis of variance

Considering the analysis of data in which the outcome is continuous and the determinant is categorical, and there is also a categorical covariate, this leads to a modification of anova called two-way analysis of variance. Suppose that there are n_j observations in sample j, denoted by y_{ij} for i = 1, 2, ..., n_j. The F-statistic is

$$F = \frac{(S_2 - S_{12})/(c-1)}{S_{12}/(n-c-r+1)}$$

where $$S_2 = \sum_{j=1}^{c}\sum_{i=1}^{r}(y_{ij} - \bar{y})^2, S_{12} = \sum_{j=1}^{c}\sum_{i=1}^{r}(y_{ij} - \bar{y}_j + \bar{y})^2$$

and $$\bar{y}_i = \frac{1}{c}\sum_{j=1}^{c} y_{ij}, \bar{y}_j = \frac{1}{r}\sum_{i=1}^{r} y_{ij}, \bar{y} = \frac{1}{rc}\sum_{j=1}^{c}\sum_{i=1}^{r} y_{ij}$$

S_2 is the sum of squares of the data after adjusting for row effects, S_{12} is the sum of squares after adjusting for both row effects and column effects. The p-value is the area in the tail of the F-distribution with $c-1$ and $n-r-c+1$ degrees of freedom (McNeil, 1996: pp 73).

Multiple linear regression

Regression used to analyse data in which both the determinants and the outcome are continuous variables. It can summarise the data in the scatter plot by fitting a straight line. In conventional statistical analysis the line fitted is the *least squares line*, which minimises the distances of the points to the line, measured in the vertical direction. In If there is more than one determinant, the method generalises to multiple linear regression, in which the *regression line* extends to the multiple linear relation represented as (Mcneil, 1998 : 185)

$$Y = \beta_0 + \sum \beta_i x_i + \varepsilon ,$$

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).

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

Reliability

A reliable measuring instrument behaves similarly: the test yields similar results when different people administer it and when alternative forms are used. When conditions for making the measurement change, the results of the test should not be used.

One of the most commonly used reliability coefficients is Cronbach's alpha. Alpha (or α) measures the "internal consistency" of a test. That is, it is based on the average correlation of items within the test, if the items are standardized to a standard deviation of 1, or on the average covariance among items on a scale, if the items are not standardized. We assume that the items on a scale are positively correlated with each other because they are measuring, to a certain extent, a common entity. If items are not positively correlated with each other, we have no reason to believe that they are correlated with other possible items we may have selected. In this case, we do not expect to see a positive relationship between this test and other similar tests.(Mariji, 1993)

Correlation Coefficient

The correlation coefficient is a measure of the linear or straight-line, relationship between variables and level of relation. The model of correlation coefficient is defined as (McNeil, et al, 1998: 181)

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

It may be shown that r ranges from a minimum of -1 to maximum value of 1. A correlation coefficient equal to 0 indicates no linear relationship between the two variables.