CHAPTER 3

METHODOLOGY

3.1 Introduction

This chapter discusses the methodology used to obtain and analyse data. Section 3.2 details the methodology used in the survey that was conducted in 1989, which is the main source of data for this study. Data on factors influencing career aspirations is obtained and this data is used to extract information on the dependent and independent variables, which are detailed in Section 3.3 and 3.4.

The analyses used in the study are bivariate and multivariate analyses. Bivariate analysis in terms of exploratory data analysis consists of graphical displays and in terms of confirmatory data analysis consists of t tests and ANOVA tests. This information is in Section 3.5. Section 3.6 outlines the method used for the multiple regression model. Section 3.6.1 and Section 3.6.2 consists of the assumptions of the model and steps taken to detect violations of these assumptions plus steps to reduce or remedy these violations. Section 3.7 provides the method used for the selection of the self concept independent variables. Then in Section 3.8, diagnostic measures in terms of residual analysis are administered to detect outliers and influential observations. Section 3.9 discusses the factor analysis statistical tool which is supplementary to the multiple regression model. Factor analysis is used to obtain a reduced number of factors representing perceived work values which are used as independent variables influencing career aspiration. Finally the computer packages used for the analysis is reported in Section 3.10.

3.2 Data Source

3.2.1 Survey

The main source of data is obtained from a survey conducted in 1989 for the University of Malaya's tracer study on "Transition from School to Work" (see Chew et al., 1995 for details). This survey is the first part of a longitudinal study and was conducted on two samples, consisting of the Form 5 students and the Form 6 students. Details concerning the socio-economic background of respondents, their academic achievements and their educational and career aspirations were captured in this survey (see Appendix 1 for the questionnaire used for data collection).

Below is a brief summary on the methodology used in the survey.

a) Sampling frame

The sampling frame of the study consists of schools offering fifth and sixth form classes in four states in Malaysia, namely Kelantan, Johore, Federal Territory of Kuala Lumpur and Selangor. These states represent different levels of economic development. Kelantan represents the low-income state, Johore, the middle-income state, and the Federal Territory of Kuala Lumpur and Selangor represent the high-income states.

b) Sample and Sampling Methodology

Two samples were obtained from this survey, one consisting of Form 5 students and the other consisting of Form 6 students. The survey yielded samples of 7944 Form 5 students and 3283 Form 6 students.

The sampling method used to obtain the samples was the disproportionate stratified sampling method. The sampling frames were stratified according to state, urban-rural locality, stream of study (academic, technical, vocational and residential streams) and school size. Within each stratified group a random sample was chosen. As disproportionate sampling was used, it was therefore necessary to assign weights to the samples. These weights were appropriately assigned to each of the observations in the

ample. The weighted samples consist of 8,141 Form 5 respondents and 2,867 Form 6 espondents.

Detailed information on the sampling frame, sampling method and sample size :an be obtained from Chew et al. (1995).

).2.2 Additional Sources

\dditional information on the socio-economic framework of the country is obtained from :conomic and Statistical Reports. Data obtained from these secondary sources is mainly)n educational policies, educational institutions, financing of education and on the labour narket and economic situation in the country.

3.3 Dependent Variables

The dependent variables are the idealistic and realistic components of career aspiration :ermed as career preference and career expectation. These two components are examined for both the Form 5 and the Form 6 samples. When the term 'career aspirations' is used, it includes both, career preference and career expectation.

3.3.1 Trieman's Standard International Occupational Prestige Scale

The career preference and expectation in this present study were initially categorised according to the International Standard Classification of Occupations (ISCO) scale but were later converted on the Standard International Occupational Prestige (SIOP) scale to obtain the prestige level for each occupation (Trieman, 1977). The prestige points obtained are within the range of 14 to 78.

Reason For Use Of The SIOP Scale

The reason for the use of this scale is as follows:

- the scale is taken to be a valid representative of the prestige hierarchy for most countries. Trieman's (1977) study showed that many countries, especially

industrialised countries had prestige data which were highly correlated with the SIOP scale. Hence it is assumed that Malaysia too, as a developing country, will have an occupational prestige hierarchy which will be highly correlated with this scale.

- The ISCO scale is not used in the analysis because the categories in the ISCO scale does not capture the differences in the prestige of an occupation. In the ISCO scale, occupations are categorised according to the type of work done and this produces considerable heterogeneity in terms of prestige within the categories. For instance, working proprietors in the sales and service industries were allocated to the sales and service categories with other lower prestige sales and service workers. Another example is that clerical supervisors were allocated to the same category as lower level clerical workers. The SIOP scale on the other hand is able to capture the prestige levels of individual occupations.

Assumptions Of Using The SIOP Scale

The assumptions made when using the SIOP scale in this study are as follows:

- As Trieman (1977) assumes a world wide occupational hierarchy, it is assumed that the scale can also be used in the Malaysian context.
- As the prestige hierarchy is also assumed to be invariant across time, it is assumed that the present ratings of occupations in terms of prestige is similar to the ratings in 1977 when Trieman's (1977) scale was constructed.
- Trieman (1977) also assumes that students and adults have similar prestige ratings and therefore it is assumed that the students in this study will have similar ratings of occupations as the adult ratings used in Trieman's scale.

3.3.2 Categorisation Of Occupational Prestige

Approximate cutting points for the prestige levels are determined to distinguish between what is considered to be low, medium, high and very high prestige careers.

1. Low Prestige Careers (Less or equal to 46 points)

2. Medium Prestige Careers (47 to 57 points)

3. High Prestige Careers (58 to 66 points)

4. Very High Prestige Careers (More or equal to 67 points)

To obtain these categories, the cutting point between high and above and medium and below prestige is first obtained. Trieman's (1977) cutting point between high and low professional and technical occupations is 58 points. This point is taken in this study to be the point that will distinguish between high and above and medium and below prestige careers. This is because in this study, the respondents' career preferences seem to be centred around professional and technical careers (64.7% for the Form 5 career preference and 80.5% for the Form 6 career preference). The 58 cutting point is also the mean prestige point for the Form 5 respondents' career preference and is close to the mean prestige of Form 6 respondents' career preference. Career preference is taken more into consideration when determining the cutting points and not career expectation as career preference reflects more the perception of respondents on occupations that are considered prestigious. Trieman's scale too is based on studies of the perception of respondents on what they consider to be prestigious occupations. This leads to more emphasis being placed on formulating categories that will be able to discriminate the career preference of respondents. If the 58 cutting point is taken, the careers that will lie in the high and above prestige categories will also include general administrative and managerial positions and administrative and managerial positions in the production, agriculture, sales and services sectors..

Next, 67 points is taken to be the cutting point to discriminate between occupations that are considered to be of very high prestige and those of just high prestige. This point is the mean prestige point for this study of occupations which are of high prestige and above (more than 58 points). Trieman's definition of occupations of very high prestige are those with 80 points and above. This point is not feasible to be used in this study as there would be no respondents in this category as the respondents do not have aspirations of that great a height.

Last of all, the mean prestige of Form 5 career preference of 46 points, which is slightly lower that the Form 6 mean prestige is taken as the cutting point to discriminate between occupations which are of medium prestige and those of low prestige. Although Trieman's (1977) study suggests that occupations of 10 points and below should be considered that of very low prestige, this point too is not feasible in this study as there would be no respondents in this category.

3.4 Independent Variables

Table 3.1 gives a description of the variables that are expected to influence career aspirations. The expected influence of some of these variables is shown in Table 2.1 of Chapter 2.

The self concept variables and its interactions are the most important variables in this study and therefore their bivariate as well as their multivariate influence are examined. The self concept variables are categorised into four groups: i) demographic variables; ii) family characteristics; iii) socio-economic status; and iv) academic related influence

Perceived work values are expected to have only a slight influence on career aspirations of adolescents so only their multivariate influence, after including the self concept and interaction variables, are examined. Career preference is expected to have an influence on career expectation and this is examined in the Chapter 6 on multivariate analysis.

Most of the variables are adequately explained in Table 3.1. However, a few variables need further explanation.

Parental Interest

Parental interest is the perception of the respondents on how much interest is shown by their parents on their welfare. It is proxied by the sum of the following responses:

- 'Very important' for the question on how important it is to the parent or guardian that the respondent does well in the examination;

'Tertiary education' for the question on highest level of education that the parent or guardian expects from the respondent;

'Very often' for the question on how often the parent or guardian checks the respondent's homework; and

Any response other than the response that the parents are not bothered, for the question on the parents' reaction if the respondent achieved good results and the question on parents' reaction if they did not achieve good results.

For regression analysis, other than the sum of responses to questions on parental ıterest, the individual responses are also tested for significance to see which haracteristics of parental interest significantly influences career aspirations.

'arents' or Guardian's Occupation and Income

For respondents whose guardian is either both parents or a single father, only the ccupation and income of the father is taken into consideration. This is in accordance vith the father's role as the breadwinner of the family. For respondents whose guardian is . single mother, only the occupation and income of the mother is taken into onsideration. Last of all, if the respondent is staying with a guardian other than the ıarent, the occupation and income of that guardian is taken into consideration.

Categorical Variables

For bivariate analysis, all the categories for the categorical variables shown in the Table 3.1 are used in the analysis. For the multivariate analysis, the categorical variables ıre converted into dummy variables of 1 and 0. The categories given the value 1 and 0 for ach variable will be shown later in Chapter 6 on multivariate analysis.

Table 3.1: Description of Variables Considered for Data Analysis

/ariable Description	Variable Type	Categories in Variable	Questionnaire No (data source)
ɜelf concept Variables			
. Demographic Variables			
ɜex	Categorical	female male	1.2
ɜthnicity	Categorical	Malay / Bumiputera Chinese Indian Other ethnic groups	1.3
ɜocality as in place that ɜespondent grew up in.	Categorical	Rural Small Town Large Town	1.13
ɜ. Familay Characteristics/ Family Influence			
ɜamily size	Continuous (for Multivariate analysis) Categorical (for Bivariate Analysis)	For Categorical Variables 1) 1-2 children 2) 3-4 children 3) more than 5 children	1.5
Birth order	Continuous (for Multivariate analysis) Categorical (for Bivariate Analysis)	For Categorical Variables Eldest / 1st child 2nd child 3rd or 4th child 5th child	1.6
Parental Interest	Continuous (for Multivariate analysis) Categorical (for Bivariate Analysis)	For Categorical Variables Each category represents the sum of responses to questions on parental interest	2.10 - 2.14
3. Socio-economic Status			
Father's highest level of education	Categorical	No Formal Education Primary Lower Secondary Upper Secondary College or University	1.8
Mother's highest level of education	Categorical	No Formal Education Primary Lower Secondary Upper Secondary College or University	1.8
Parents' or Guardian's Occupational Status	Categorical	Low Prestige Medium Prestige High Prestige Very High Prestige	1.9
Parents' or Guardian's Income	Categorical	RM0-RM500 RM501-RM1000 RM1001-RM2000 More than RM2000	1.11

4. Academic Related Influence			
Academic achievement *SRP* (for Form 5) *SPM* (for Form 6)	Continuous (for Multivariate analysis) Categorical (for Bivariate Analysis)	For bivariate analysis, the *SRP* variable is recoded into quartiles and the *SPM* variable is recoded into quintiles.	2.3 or 2.5
Stream of Study	Categorical	Science Arts Others	2.1
School Leaving Decision Making	Categorical	Further education immediately Work for a while before pursuing further education Seek employment immediately	2.21
Interaction Variables	Interaction of gender and ethnicity with all the socio-economic status variables. Interaction of ethnicity with academic achievement		
Career preference as independent variable	Only influences career expectation		
Work Values	Factor scores used (will be explained in more detail in section 3.9)		3.5

.5 Methodology Used In Bivariate Analysis

wo types of analysis is used in the bivariate analysis, that is, the exploratory data ıalysis consisting of graphical displays of the data using box plots and confirmatory ıta analysis consisting of tests for mean differences in career aspiration for the different ıtegories.

.5.1 Exploratory Data Analysis

ıox-plots

To compare differences in distribution of different groups of cases, recommended ıethods to use are either a dot-plot or box-plot (Tukey,1977). A box-plot was used in ıis study as it is easily obtained using the SPSS programme.

The box-plot (Tukey, 1977) represents an abbreviated histogram. The rectangle in ıe box-plot extends from the lower quartile to the upper quartile of the sample data, and ; divided at the median. The horizontal lines extend from the ends of the rectangle as far s the extreme data values within a distance equal to the interquartile range. Points at the nds denote more extreme data values known as 'outliers'.

By comparing box-plots, it is easy to eye-ball the amount of overlap there is in wo distributions. There is a larger effect if the interquartile range of one group does not ıverlap with the interquartile range of another.

It is also possible to see if the distribution is skewed by checking the distances ıetween the ends of the horizontal lines (extreme values) and the median or the distances ıetween the upper and lower quartiles and the median. If the upper and lower distances ıre equal, the distribution is symmetrical. If the upper distance is longer, the distribution s positively skewed. If the horizontal lines are longer for a particular group, it indicates hat the data is more widely spread.

3.5.2 Confirmatory Data Analysis

To estimate the mean differences between the categories of independent variables, the t test or F test is used. The t test is used to compare the means of two categories and the F test is used to compare the differences in mean when the number of categories is more than two.

a) Independent Sample t Test

The independent sample t test compares the mean for two groups of cases.

Ho: $\mu_1 = \mu_2$
Ha: $\mu_1 \neq \mu_2$

Test statistics

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

Where: $v = n_1 + n_2 - 2$, $s^2_p = \frac{(n_1 - 1)s^2_1 + (n_2 - 1)s^2_2}{n_1 + n_2 - 2}$

Critical Region: $t < -t_{\alpha/2}$ and $t > t_{\alpha/2}$

Where:

s_p	=	pooled variance
v_1 and v_2		degree of freedom for category 1 and 2
μ_1 & μ_2	=	population mean for category 1 and 2
$\bar{x}_1$ & $\bar{x}_2$	=	sample mean for category 1 and 2
n_1 & n_2	=	sample size for category 1 and 2

the significance level, α is taken to be at the 5% level

If the null hypothesis is rejected, we can be 95% confident that the mean of the two categories are significantly different

Assumptions of the independent sample t test:

- Independent samples
- Normal population

- Variance of the two groups are equal

b) One Way Analysis of Variance (ANOVA)

The ANOVA test (refer to Kleinbaum and Kupper, 1977) compares the mean for two or more groups of cases.

Ho: $\mu_1 = \mu_2 = \ldots\ldots\ldots = \mu_k$
Ha: the means of at least some of the categories are not equal

Test Statistics: $F = \dfrac{MST}{MSE}$

Where:

$$MST = \left[\sum_{i=1}^{k}\left(\frac{T^2_i}{n_i}\right) - \frac{G^2}{n}\right] \Big/ k-1$$

$$MSE = \left[\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y^2_{ij} - \sum_{i=1}^{k}\left(\frac{T^2_i}{n_i}\right)\right] \Big/ n-k$$

Where:

μ_i = population mean of category i; T = Total for each category; G = Grand Total; k = number of categories; n = sample size;

Y_{ij} = observation, MSE = Mean squared error, MST = Mean squared treatment

Critical Region: $F \geq F_{(k-1)(n-k)(1-\alpha)}$

Where the significance level, α is taken to be at the 5% level

If the null hypothesis is rejected, it can be concluded at the 95% level that the means of some of the categories are not equal.

Assumptions of ANOVA are as follows:

- Independent samples

- Normally distributed population
- Equal population standard deviations

c) Multiple Comparisons of Mean - Scheffe Test

Sheffe test enables multiple comparisons of the means (refer to Kleinbaum and Kupper, 1977). For this test, the probability is (1-α) that all possible differences in mean are in a set of intervals given by:

$$|\overline{x_a} - \overline{x_b}| \pm s\sqrt{MSE\left(\frac{1}{n_a} + \frac{1}{n_b}\right)}$$

Where: $s = \sqrt{(k-1)F_{k-1,n-k,1-\alpha}}$

Ho: $\mu_a = \mu_b$
Ha: $\mu_a \neq \mu_b$

Reject the null hypothesis if $|\overline{x_a} - \overline{x_b}| > s\sqrt{MSE\left(1/n_i + 1/n_j\right)}$.

Where:	MSE	=	Mean squared error
	μ_a	=	population mean for category a
	$\overline{x_a}$	=	sample mean for category a
	n_a	=	sample size for category a
	k	=	number of categories

α is significant at the 5% level

When the null hypothesis is rejected, it can be concluded at the 95% confidence level that there is a significant difference in mean between category a and category b.

<u>Reasons for the use of the Scheffe test instead of other tests of multiple comparisons of means such as the Tukey test is as follows:</u>

- Sheffe test is recommended to be used when the size of the different samples are not equal. In this study, the categories for most of the variables are not equal in size.

- Even when the assumptions of normality and equal population variances are not satisfied this test can still be used. The categories for some of the variables in this study may not have equal variances.

- When the F test rejects the null hypothesis that the means of all the categories are equal, the Scheffe test will in most cases also reject the null hypothesis at least once to show that at least one of the pairs of means are not equal. This is not so for the Tukey test where at times the F test may reject the null hypothesis, but the Tukey test may still show that all the means of the categories are equal to one another.

d) Paired Sample t test

This test is used in the section on sample characteristics to see if there is a difference in the mean of career preference and career expectation. This test is used as career preference is found to be a precursor to career expectation and therefore they are considered as paired groups.

<u>Assumptions of the Paired t Test</u>

- Paired samples

- Population of paired differences are normally distributed

Ho: $\mu_D = 0$
Ha: $\mu_D \neq 0$

Where μ_D is the mean of the paired differences

$d = x_1 - x_2$

Test statistics: $t = \dfrac{\bar{d}}{s_d / \sqrt{n}}$

Where s_D is the standard deviation of the paired differences

<u>Critical region</u>: $t < -t_{\alpha/2}$ and $t > -t_{\alpha/2}$

Where: α is the significance level, taken to be at the 5% level

If the null hypothesis is rejected, we can conclude at the 5% significance level that there is a significance difference in the means of the paired samples.

3.6 Multiple Regression Analysis

The main objective of the multiple regression models will be to determine the variables that significantly influence career aspiration and the relationship between these variables and career aspiration. The net effects of each variable on career aspiration can be determined. As the objective is not to predict career aspiration, emphasis will not be placed on achieving maximum predictive power (R^2) but instead, on finding important determinants of career aspiration.

In view of the fact that career aspiration consist of career preference and career expectation, four models are estimated to explain career aspiration, that is one for each dependent variable and one for each group, the Form 5 sample and the Form 6 sample.

3.6.1 Classical Assumptions Of Multiple Regression Model

The regression model is:

$$Y_i = \beta_0 + \beta_1 X_1 + \ldots\ldots\ldots\ldots \beta_p X_p + \mu_i$$

Assumptions of the multiple regression model as stated in Gujerati (1995), is as follows:

- Zero mean values of μ_i or $E(\mu_i \mid X_1 \ldots\ldots X_p) = 0$
- No serial correlation or $\text{cov}(\mu_i \mu_j) = 0$ $\quad i \neq j$
- Homoscedasticity or $Var(\mu_i) = \sigma^2$
- Zero covariance between μ_i and each X variable, or $\text{cov}(\mu_i X_1) = \text{cov}(\mu_i X_2) \ldots = 0$
- Reduced correlation among independent variables, i.e. multicollinearity is low

3.6.2 Violation In The Above Assumptions

a) Non-Normality Of Error Terms

The histogram of the standardised residual and also the normal probability plots of the standardised residual is used to check for normality of the distribution of the error term (Hair et al., 1998).

Normality should show a symmetrical shape (not negatively or positively skewed) in the distribution and the shape should not have the characteristics of kurtosis (referring to peakedness or flatness of the distribution).

The normal probability plots of the standardised residuals compare the cumulative distribution of the actual data with the cumulative distribution of the normal distribution. The normal distribution makes a straight diagonal line, and the plotted residuals are compared with the diagonal. If a distribution is normal, the residual plots will closely follows the diagonal.

b) Multicollinearity

If there is multicollinearity among variables, then it would be hard to interpret the influence of a variable on career aspiration because the coefficient consist of the unique effect of that variable and the shared effect of other variables. This section shows the attempt made to minimise multicollinearity.

Detection of Multicollinearity

The following are some signs or methods used in this study to detect multicollinearity (refer to Gujerati,1995):

1. Preliminary analysis of the correlation matrix to check for high pair-wise correlation among the variables.

2. A priori information indicating that there may be high correlation among some variables.

3. Low values for tolerance level (TOL) or high values of Variance Inflation Factor (VIF) may be an indication of multicollinearity. Variance Inflation Factor (VIF) shows how much the variance of an estimator is inflated by the presence of multicollinearity.

$$VIF = \frac{1}{1 - R_j^2} \qquad \text{and} \qquad TOL = \frac{1}{VIF}$$

where R_j^2 is the R^2 in the (auxiliary) regression of the regressor X_j on the remaining regressors.

Removing and adding different variables at a time and observing its effect on the TOL and VIF will give an indication on which variables are correlated with each other.

4. Cross tabulating some variables also gives an indication on whether correlation among variables exist.

Measures to reduce multicollinearity

Multicollinearity is reduced by dropping a few variables from the model. Variables that are found to be important in the determination of career aspiration will however not be dropped from the study because the objective of the study is to find important determinants of career aspiration.

Some guidelines are used to decide on which variables to drop. These guidelines are used based on the fact that they may ensure that more important variables are included in the model and the less important ones are dropped. The guidelines are as follows:

1. The coefficients and t values of the independent variables that are found to be correlated in the model are considered. The ones with lower coefficients and t values are considered to be removed from the model.

2. The adjusted R^2 of reduced models after dropping a variable and t values and coefficients of the variables that were not dropped are considered. A higher adjusted R^2 for the reduced model with higher coefficients and t values for the variables that were not dropped indicates that the reduced model is a better one.

3. Partial correlation of the correlated independent variables are considered. Variables that have a lower partial correlation with the dependent variable is considered to be removed from the model.

c) Heteroscedasticity

Heteroscedasticity is a scenario where the variance of the residual is not equal for all observations. This is a violation of the multiple regression assumption where the Ordinary Least Squared (OLS) method assumes equal variance i.e., homoscedasticity.

Detection of heteroscedasticity

Heteroscedasticity is detected using the graphical method and the formal method such as the Park test and the Spearman's rank correlation test (refer to Gujerati, 1995).

1) Graphical method

First the estimated residual squared, $\hat{\mu}_i^2$ of the regression models are plotted against the estimated dependent variable, $\hat{Y}_i$. If a pattern can be observed in the graphs, then there is some indication that heteroscedasticity may exist in the model.

The estimated residual squared, $\hat{\mu}_i^2$ of the regression models are also plotted against one of the explanatory variables. If the graphs portray a pattern, then this suggests that the heteroscedasticity variance may be proportional to the value of the explanatory variable.

The explanatory variables that will be considered are mainly academic related variables. This is in line with the theory that higher prestige careers require people of higher educational levels. Hence people who are highly educated have the choice of high prestige careers as well as low prestige careers. It is assumed that respondents who have high academic achievement and intentions will also realise that they have a wider choice

and therefore the variance in the distribution of career aspiration is expected to vary according to different academic related characteristics of respondents.

2) Park Test

If from the graphical analysis, there is some indication of heteroscedasticity, then the Park test is performed (refer to Gujerati, 1995)

Firstly, the following regression is run:

$$\ln \hat{\mu}_i^2 = \alpha + \beta \ln X_i + v_i$$

where $\hat{\mu}_i^2$ is the estimated residual squared and X_i is the explanatory variable

If β turns out to be significant, it would suggest that heteroscedasticity is present in the model.

3) Spearman's rank correlation test

Speaman's rank correlation is also used to test for heteroscedasticity in the model (refer to Gujerati, 1995). The following steps are taken to run this test:

(i) The estimated residual squared, $\hat{\mu}_i^2$ is first obtained.

(ii) The absolute value of $\hat{\mu}_i^2$ and the value of the explanatory variable, X_i are ranked according to an ascending order

(iii) The Spearman's rank correlation is then computed as follows:

$$r_s = 1 - 6\left[\frac{\sum d_i^2}{n(n^2-1)}\right]$$

where: d = difference in the rank assigned to the two variables,
n = number of observations

(iv) r_s is then tested using a t test as follows:

$$t = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}} \qquad \text{with df} = n-2$$

If the t value exceeds the critical t value, then the hypothesis that heteroscedasticity exists is accepted., otherwise it is rejected.

Remedial Measures

If heteroscedasticity exists in the model and if it is found that the heteroscedasticity variance is proportional to an explanatory variable in the model, then the weighted estimation procedure is used. For this procedure, the coefficients of the regression models are calculated using the weighted least square method. The explanatory variable that is found to be proportional to the heteroscedasticity variance is used as the weight variable and the best possible transformation which fits the data best is obtained using the weight estimation procedure obtained from the SPSS computer programme.

3.7 Strategy For Selection Of Self Concept Independent Variables

The self concept variables that are selected as independent variables are based on the following criteria:

- They are expected to be important variables based on theory and past studies.
- When included in the regression models, they are significant at the 5% or close to 5% levels or if the variable have signs that are interpretable and concur with the theory
- If they do not correlate highly with any of the other variables in the regression model i.e. no multicollinearity effect.

This section will concentrate on the first two criteria. Multicolinearity has already been explained in Section 3.6.2.

1) Firstly all the self concept variables and the interaction terms (shown in Table 3.1), which are expected to be the most important variables influencing career aspirations, are included in the models. The categorical variables are converted into dummy variables with two values, 1 and 0 (as base category). The way to split into these two

categories is by identifying via bivariate analysis, where the main differences in career aspiration lie

Variables are removed one by one and their effect on the regression model is observed. To determine what variables to remove from the model, the following guidelines are used:

i) If the t tests show that the variable is insignificant or if the variable has a sign that is not interpretable or does not concur with the theory, then the variable is considered for removal. The significance level is taken to be at the 5% or close to 5%. level

ii) If when the model is re-estimated after removing a particular variable, the adjusted R^2 increases or is constant, than the variable is removed from the model. A higher or constant adjusted R^2 shows that the exclusion of the variable increases or does not make a difference to the explanation power of the model. However, if the adjusted R^2 is much lower, than the variable is not removed as the variable is considered an important variable in explaining the dependent variable.

iii) If (i) and (ii) results in the decision to remove the variable, the variable is not removed immediately. Other insignificant variables which may be related to this variable is removed first from the model. The removal of this other variable may cause the previous variable to be significant. For instance, some socio-economic status variables may have an influence on the other socio-economic status variables. Therefore, if parents' income and parents' occupational status are insignificant, it will first have to be ascertained whether the removal of one will cause the other to be significant. Then, the variable that contributes most to the regression model (in terms of highest R^2 for the reduced model) is left to remain in the model.

iv) For dummy variables which are insignificant, a different method of categorisation is used to see if it makes a difference to the significance of the variable.

3.8 Diagnostics

3.8.1 Outliers

Outliers are observations that have large residual values that can be identified with respect to a specific regression model. Outliers are identified by examining the residuals (refer to Hair et al., 1998). The studentized residual, as a primary indicator of outliers, is examined to identify points that are outside the upper and lower limits which is taken to be -2 and 2 (this is the threshold level mentioned in Hair et al., 1998)

After the outliers have been identified, they are temporary ommitted, and the regression model is re-estimated. This is to see whether outliers have an influence on any of the coefficients of the regression models.

3.8.2 Single Case Diagnostics: Identifying Influential Observations

The influence of a single observation on the results is also examined. The indicator of influential observations used in this study is the DFBETA (refer to Hair et al., 1998). DFBETA is calculated as the change in the coefficients when an observation is deleted. It is the relative effect of an observation on each coefficient. Points that fall outside the upper and lower limits which is taken to be $[\geq (\pm 2 \div \sqrt{n})]$ (this is the threshold level mentioned in Hair et al., 1998), indicate that the observation is an influential observation.

DFBETA is used in this study and not other indicators as the purpose of this study is to find important variables influencing career aspiration, and by using the DFBETA, it can be observed whether certain variables are significantly influencing career aspirations only because of the influence of influential observations.

Influential observations for some variables are temporarily omitted and the models are re-estimated. This is to check whether the independent variable influenced by these observations still have a significant influence on career aspirations after omitting the observations.

3.9 Factor Analysis

3.9.1 Objective And Design

The objective for using factor analysis in this study is to identify representative variables for perceived work values from a larger set of variables, to be subsequently used in the multivariate analysis (refer to Hair et al., 1998).

The perceived work value variables are measured on an ordinal scale of 1 to 4 with one representing very strong importance placed on a particular work value and the least importance is coded as 4.

A factor analysis is performed on the perception of respondents on the importance of different work values for job satisfaction. This is to reduce it to a smaller number of interpretable factors. Factor analysis is also conducted on work values that are perceived to be important motivators or barricades for working in rural areas. The two factor analyses are conducted separately because it is presumed that the factors affecting job satisfaction and those affecting the motivators and barricades for working in rural areas are different. The principle component method is used when extracting the factors.

3.9.2 Assumptions

(i) Normality

(ii) Homoscedasticity

(iii) Linearity

Assumptions (i), (ii) and (iii) are also assumptions made for the multiple regression model.

(iv) Data matrix has sufficient correlation

The variables are examined to see if there is sufficient correlation among variables and that the sample is adequate for conducting a factor analysis. This is done by examining the correlation matrix, the Kaiser-Meyer-Olkin measure of sampling adequacy

and the Bartlett's test of sphericity. The correlation matrix is inspected to ensure there is sufficient number of correlations of greater than 0.3 (rule of thumb as stated by Hair et al., 1998). The Kaiser-Meyer-Olkin measures the degree of inter-correlations among the variables and a measure of around 0.8 and above is graded as good (as stated in Hair et al., 1998). The Bartlett test of sphericity provides the statistical probability that the correlation matrix has sufficient correlation among the variables. If the null hypothesis for this test is rejected, than it can be concluded that there is sufficient non-zero correlation at a given significance level.

Variables that do not correlate highly (correlation coefficient less than 0.3) with any of the other variables is removed from the factor analysis and will be used separately as an independent variable.

3.9.3 Selecting A Factor Method

The method for factor extraction is the principle component method. The principle component method is used as it is appropriate when the primary objective is to minimise the number of factors needed to account for the maximum portion of the variance represented in the original set of total variables. Common factor is not selected as it poses many problems such as no single solution, substantial computer-time and resources plus the communalities found are not always estimable or are sometimes invalid leading to the deletion of certain variables from the analysis (Hair et al., 1998).

3.9.4 Extraction And Interpretation Of Factors

To determine the number of factors to be extracted, factors having eigenvalues more than 1 are first considered. The factors are then rotated using the varimax rotation method (Hair et al., 1998). By examining the factor loadings of variables and the combinations of variables loading heavily on each factor, it can be determined whether these factors are practically interpretable. Super(1970)'s work values (see Appendix2) are used as an aid for interpreting the factors. If the number of factors in the first step are deemed not suitable for interpretation and also if the cumulative percentage of variance extracted by the factors is not sufficient (at least more than 60%), the factor extraction is repeated a

v times by also considering factors having eigenvalues close to 1 such as 0.8 or 0.9. e final number of factors extracted will consist of rotated factors that have eigenvalues se to 1, have cumulative percentage of variance extracted more than 60% and are erpretable in terms of the combination of variables loading heavily on each factor air et al., 1998). These factors are then named by taking into consideration terminology ed in Super's (1970) Work Value Theory (see Appendix 2).

).5 Computation Of The Factor Scores

rceived work value factors, in terms of factor scores are then included in the models to e if they have any influence on career aspirations.

It should be noted that perceived work values that are rated as most important are ven a rating of 1 and least important is rated as 4. Hence, the higher the factor score, the wer the importance placed on work values and a negative coefficient for work values dicates a positive influence between work values and career aspirations.

10 Computer Packages Used

ll computations are done via the SPSS (Statistical Package for Social Science). ogramme. Figures such as histograms and box plots are also obtained using this atistical package.