Chapter 3 Methodology

3.1 Introduction

This chapter discusses the methodology used in this study. Section 3.2 discusses the methods used to measure mobility while Section 3.3 consists of a brief introduction to International Socioeconomic Index (ISEI). Section 3.4 gives a brief discussion about the data sources and Section 3.5 contains discussion about the dependent variables studied. Section 3.6 provides the list of independent variables used. Section 3.7 and Section 3.8 give the information about the computer programmes used and a summary of the statistical procedures used in this study respectively.

3.2 How is mobility measured?

This section is divided into two parts. The first discusses approaches to measuring mobility and the second the scales used to measure occupational status and income level.

3.2.1 Approaches to measuring mobility

Khandker (1992) operationalised occupational mobility as the number of times a worker changes job divided by the length of time he or she works.

Li (1977) measured occupational mobility by treating the first job as an origin's status and the current job as the destination status. The question of mobility involves evaluating the effect of the origin status on the destination status (Blau and Duncan, 1967). The regression coefficient is used to measure the effect of the origin status on the destination status, i.e. one unit change in origin destination causes how much change in the destination status.

Blau and Duncan (1967) found that the mobility score (the difference of occupational score between the origin and destination location) is not commonly acceptable and statistical analysis using mobility score may incur confounding effect. This is due to the fact that mobility phenomenon is not homogeneous and the factors determining where an individual starts may be different from the factors determining the destination status he achieves. Furthermore, mobility to certain occupations may be easier from one point than from another point. In other words, the origin score is one of the factors determining the destination point.

Blau and Duncan (1967) nevertheless examined the mobility score distribution for two main reasons. Firstly, it is interesting to know which population group enjoys more or less amount of mobility. Secondly, it gives interesting and valid characteristics of the mobility process and serves as a way to check the conclusion reached by other means.

Leigh (1978) used mobility scores as dependent variable and the initial occupational status as independent variable on the right hand side. The example below is to give a closer look of the mobility score model. Suppose, education is positively related to the first occupational status and first occupation is negatively related to the mobility score. If the first occupational status is omitted, the coefficient for education understates the true impact of education on occupational advancement. The first occupation variable is used to measure the initial occupation standing so that the education variable only evaluates the effect of education on occupation advancement. The absence of the first occupation status variable would imply that the coefficient on education confounds the positive effect of schooling on occupational mobility with the negative relationship between occupation change and first occupation level. So, it is important to include the first occupation status in the model.

Besides, Leigh (1978) also used the probability of occupational upgrading as a measure of occupational mobility. It is binary variable, which takes the value 1 if the individual enjoys upward mobility and 0 otherwise.

3.2.2 Scales to measure mobility

Li (1977) used prestige scale to measure occupational status. The scale used was Duncan's socioeconomic index. This scale is also used by Blau and Duncan (1967) and Leigh (1978). This measure considers both the education and income of an occupation. There is overlapping of scores assigned to different major occupational groups. Occupational structure is assumed to be continuously graded in regard to status. The justification is as below. Occupational groups overlap in a greater or lesser degree, in terms of income, education attainment, consumer expenditures, measured intelligence, political orientations, residential locations and others. So, it is reasonable to treat occupation structure as continuous.

The disadvantage of the prestige scale is that it has a built-in positive correlation with education. Hall and Kasten (1976) suggested that SES has a bias against blue-collar jobs (quoted in Leigh, 1978).

Chattopadhyay (1998) used the International Socioeconomic Index scores (ISEI) as an instrument to measure occupational status. ISEI index shows the socioeconomic characteristics of individuals in an occupational category. An individual's socioeconomic status does not change as long as he stays in the occupation.

Ornstein (1976) used two rankings to measure a given job, i.e. occupational prestige score and the wages paid to the respondent. The prestige score used is based on a ranking of over 300 occupations reported by Siegel (1970). This prestige score 'only' takes into consideration the general overview of the social standing of the occupational group in America, which was obtained in a national sample survey. In certain cases, it ranks white-collar jobs very closely to certain skilled blue-collar occupations and service occupations are scattered at different levels. It is used as a continuous prestige ranking.

For the second ranking scheme, the wage values used are the wages paid to the individual respondent. These wage values are registered at the start and at the end of each job. All the wage values are standardized to monthly income and were adjusted to yield rates of pay in constant dollars (price adjusted, to a 1959 base). These standardized value is used to compare the wage rates earned by individuals who entered the labour force during the 15-year period studied.

Leigh (1978) also measured occupational mobility by change in the income earned. The income that is used for every occupation title is the median income for male members in the occupational group. The members must be 16 years old or above.

The use of median income may understate or overstate the prestige generally associated with an occupational group. This may be due to 'compensating wage differentials'. Certain individuals are willing to accept jobs that are lower rank if the jobs have higher income. This is a trade-off between pecuniary and nonpecuniary aspects of a job.

Rose (1999) defined mobility as the absolute change in real income or earnings over ten years. To assess change over time, a three-year average income is computed.

Only individuals of 'prime' workforce age (22 – 46 years old at the start) are examined, because they are expected to be economically active throughout the period. Besides, adjusted income based on family size is used because it requires more income to support more people.

In this study, occupation mobility is defined as the change of occupation from the first full time occupation to the current full time occupation. The first full time job refers to the first full time job an individual holds after he finishes his formal education, including tertiary education, if applicable. The definition of full time job is based on the responses of the respondents. No further information was obtained such as number of hours worked in a week, which is often used to determine whether an individual is working full time or not. The measurement of occupation is based on the ISEI scale, which is discussed in greater detail in the next section.

In this study, income mobility is a direct measure of the difference between the starting monthly income for the first full time job and the average monthly income of the current job[1]. The average monthly income is inclusive of over time pay, commission, allowance and bonus. This study covered seven years and the consumer price index was relatively stable in the period under study, with the annual change ranging from 2.8% to 4.7% (Bank Negara Malaysia, 1992, 1993, 1998). Thus, although the respondents may have entered the labour market at different time points, the differences in the value of money at different entry time points within the seven years are likely to be small. The nominal income measure is therefore a reasonably adequate measure of income mobility

---

[1] The current monthly income for current job is not used as the information is not asked in the third stage questionnaire. However, average monthly income for current job is used as it is available in the data set used.

for all respondents. There is also no adjustment for family income, as suggested by Rose (1999, see the same section), given only a small proportion (12%) of the respondents have working spouses.

3.3 International Socioeconomic Index (ISEI)

The discussion on the ISEI is divided into three parts. They are definition of the ISEI, comparison of the ISEI with Treiman's Standard International Occupational Prestige Scale (SIOPS) and comparison of the ISEI with International Standard Classification of Occupations (ISCO). The discussion is based on Ganzeboom et al. (1992).

3.3.1 What is the ISEI?

The ISEI was developed as a continuous measure for occupation stratification by Ganzeboom, De Graaf and Treiman in 1992. The range of score is from 16 to 90. The ISEI is conceptualised as a measure for the characteristics of occupations that convert an individual's main resources, education, into his main reward, income. The ISEI scales occupations in such a way that it captures as much as possible the indirect effect of education on income and minimises the direct effect. The ISEI is constructed under the interpretation that occupation is an intervening mechanism between education and income. This idea is similar to that of Duncan (1961) in the construction of Duncan's Socioeconomic Index:

> "....a man qualifies himself for occupational life by obtaining an education; as a consequence of his pursuing his occupation, he obtains

income. Occupation, therefore, is the intervening activity linking income to education....." (Duncan, 1961, pp. 116-117)

By design, the ISEI is restricted to those who work full time and active in labour force. The age range is between 21 to 64 years old. The confinement to only those who work full time is to avoid the confounding effects between occupation and the status of the occupation as the status may influence the hours an individual works. This in turn may influence his income and the socioeconomic status. On the other hand, the restriction of the age is to minimise the distortion introduced by those who are going to retire and those who join the work force for a short time during the transition period from one education level to the next higher education level. One common scenario is that some people work during the period they wait for their pre-university result, before they enter universities or colleges. These two groups often have lower incomes than those employed on a regular basis.

3.3.2 ISEI and SIOPS

In this study, the ISEI is preferred compared to the Treiman's SIOPS, which is a prestige scale. There are similarities between SIOPS and ISEI, for example, both are continuous and use unidimensional approach to class occupations. However, prestige scales involves evaluation and judgement, by a sample from the population at large or a group of well-informed or experts in a society. In Treiman (1977), prestige is awarded according on the basis of power resources. In modern societies, education, considered as cultural resources, and income, a main form of economic resources, are the main power resources. Goldthorpe and Hope (1972, 1974) summarised that prestige scales measure "the general desirability of occupations" (quoted in Ganzeboom et al., 1992). In constructing the ISEI, prestige is taken as a consequence or end result, not as a parallel measure. It concerns the relationship between socioeconomic status and occupations.

Previous studies have shown that socioeconomic status of an occupation is a better representation of its status in general. Furthermore, Featherman and Hauser (1976), as quoted in Ganzeboom et al. (1992), found that "prestige score are error prone" in estimating the socioeconomic status of occupations.

One of the common problems in using an index to measure occupation status in studying occupational mobility is the problem of temporal stability. Blau and Duncan (1967) assumed *faute de mieux*. This means that the scale of occupational status remains fixed over the period studied. The same assumption is used in this study.

In using the SIOP scale, the assumption used is that the prestige hierarchy does not change over time, as given above. The SIOP was constructed in 1977, which was more than 10 years ago before the survey first started, while the ISEI was introduced in 1992. Compared to the year this survey was conducted, it seems improbable to accept the *faute de miex* assumption, if the SIOPS is chosen as a scale of measurement. The assumption is more tolerable if the ISEI is used. In addition, Ganzeboom et al. (1992) found that the ISEI compares favourably with SIOPS.

3.3.3 ISEI and ISCO

There are slight differences between the occupation list in the ISEI and the occupation list in the Dictionary of Occupational Classification, which is based on ISCO 1968 (Manpower Department, Ministry of Labour, Malaysia, 1980). Certain occupations are coded differently in the two lists. For instance, Auditors (ISCO = 111) is coded as 1109 in the ISEI and Journalist and Editors (ISCO = 152) is included in 159 category in the ISEI. On the other hand, certain occupations are not found in the ISEI occupation list. The popular example for this type of situation is the omission of different categories of managers (ISCO = 213 – 218). For this type of situation, most of them are assigned the ISEI score for the major group. For example, all managers in the above case is given the score for the Manager major group, taking the ISEI score of 67. This type of discrepancy can also be noted in other occupation categories like Statistical Clerk, Plantation Supervisor, Specialised Livestock Farmers, Rubber Processing Workers and Plastic Product Makers. All the occupation categories omitted involve a very small number of occurrences (except for managerial category). On the other hand, there are a few occupations that only appear in the ISEI, for example, Pelt Dressers and Fur Tailors. In short, the differences of occupation lists used in this survey and in the ISEI are not large enough to affect the analysis of this data severely.

3.4 The data

3.4.1 Data sources

The main data source for this study is obtained from a longitudinal survey on "Transition from School to Work" starting in 1989 (see Chew et al., 1995 for details). The three stages of the longitudinal survey were conducted in 1989, 1992 and 1996 respectively.

The first stage consisted of two samples: Form 5 and Form 6 students in selected secondary schools in 1989. Family socioeconomic background, academic achievements, attitude towards education and working, educational and career aspiration and job satisfaction were among the information collected in the first stage (see Appendix A for the questionnaire used for data collection). A total of 10927 (7944 Form 5 students and 3283 Form 6 students) respondents were included in this stage. In the second stage, only a sub-sample of respondents was selected to send the mail questionnaires.

The survey covered four states in Malaysia. Kelantan, Johore, Kuala Lumpur and Selangor were selected to depict the different level of developments. Kelantan represents the low-income states, Johore portrays the middle-income while Kuala Lumpur and Selangor sketch the high-income group picture. The sampling frame consisting two separate lists of schools offering Form 5 and Form 6 classes was used.

In the third stage, a mail questionnaire was sent to every respondents in the longitudinal survey regardless whether or not the respondents took part in the second stage survey. Details of the educational and training background, attitude towards working conditions and life values, working history and personal background information were obtained in this third stage survey (see Appendix B for a sample of the mail questionnaire). A total of 1702 respondents responded to the mail questionnaires.

This study uses the responses from the third stage. Certain demographic and educational information is extracted from the first stage. Two respondents are excluded from further analysis as no first stage information is recorded. The data set used for analysis has 1700 respondents.

A brief description on the data collection method for the first stage is presented below (Chew et al., 1995). Disproportionate stratified sampling method was used to select the samples. The sampling frames were stratified to improve the precision of sample estimates. They were stratified according to state, urban-rural location, stream offered (academic, technical, vocational and residential streams) and school size. The stratification based on gender and ethnicity was not possible. Within each stratum, a random sample was chosen.

Weighted data is used in this study as the survey in the first stage was based on a disproportionate stratified sample. The weightage in the first stage is used. Using weighted data, there is a total of 1738 respondents in this sample. It is to be noted that not all respondents answered all the questions in the questionnaires.

Besides, secondary data are also utilised to get a general and overall picture of Malaysia's economy and labour market conditions. The data are taken from Yearbook of Statistics, Economic Report, Malaysia Five-year Plan and Labour Force Survey.

3.4.2 Comparison of the first stage and third stage samples

These two data sets are compared by gender, ethnicity and SRP result. This is shown in Table 3.1. In the third stage sample, as in the sample in the first stage, the proportion of female is higher than male. There is not much difference of the percentage of each gender group in these two samples. As for the ethnic group, in the third stage sample, there is a higher proportion of Bumiputera and less of non-Bumiputera. The proportion of Indian in the third stage is about half of the proportion in the first stage. As for SRP aggregate, in general, there is a higher proportion in the lower aggregate categories, which means the respondents in the third stage generally have better result.

Table 3.1 Comparison of percentage of gender, ethnicity and SRP aggregate in first stage and third stage samples

Variable	First stage	Third stage	Variable	First stage	Third stage
Gender			**SRP aggregate**		
Female	54.3	53.2	5 – 15	36.2	42.7
Male	45.4	46.8	16 – 25	28.8	30.8
Ethnicity			26 and above	35.0	26.5
Bumiputera	63.7	69.0	Sample size	11008	1706
Chinese	28.0	26.6			
Indian	8.3	4.4			
Sample size	10979	1738			

Note: The first stage computation is based on Chew et al. (1995).

In the third stage sample, more Bumiputera and less Indian responded to the questionnaire. Using SRP result as a barometer, this sample includes respondents with better academic results. It may be the case as the better educated were more responsive to the purpose and usefulness of this survey.

3.5 Dependent variables

The dependent variables are occupational mobility and income mobility. A summary of the dependent variables is included in Table 3.2.

3.5.1 Occupational mobility score

In the longitudinal survey, all the occupation titles were initially categorised using the Dictionary of Occupational Classification (Manpower Department, Ministry of Labour and Manpower, Malaysia, 1980). It was based on ISCO 1968. So, by referring to the conversion table in Ganzeboom et al. (1992), the occupations in the survey is converted into the ISEI scales.

Occupational mobility score refers to the difference of ISEI score between the first full time occupation ISEI score and the current full time occupation ISEI score.

Upward occupational mobility refers to those who change job to a higher ISEI score occupation, compared to the ISEI score of their first occupation. In this case, the occupational mobility score is positive. Horizontal occupational mobility refers to those who change job with the same ISEI score or does not change job at all. So, the occupational mobility score is zero. Downward occupational mobility refers to those who change job to a lower ISEI score occupation, compared to the ISEI score for their first occupation. So, the occupational mobility score is negative.

3.5.2 Income mobility score

Income mobility score is the difference between the starting monthly income for the first full time job and the current average monthly income for the current full time job. It is measured in Ringgit Malaysia.

Downward income mobility occurs when an individual has a negative value income mobility score, which means a drop in income. Upward income mobility occurs when an individual has positive income mobility score, which means an increase in income. Horizontal income mobility occurs when an individual has zero income mobility score, which means no change in income.

Table 3.2 Description of the variables used

Variable Description	Variable type	Categories in variable	Question No.
Dependent variable			
Occupational mobility			
Occupational mobility scores = Current occupation ISEI score – First occupation ISEI score[1]	Continuous/ Categorical	- Negative scores (Downward mobility) - 0 (Horizontal mobility) - Positive scores (Upward mobility)	Q. 16 in 3rd stage

Income mobility			
Income mobility scores = Current occupation average monthly income – First occupation starting monthly income[2]	Continuous/ Categorical	- Negative scores (Downward mobility) - RM 0 (Horizontal mobility) - Positive scores (Upward mobility)	Q. 16 in 3rd stage
Demographic variables			
Gender	Categorical	- Female - Male	Q. 1.2 in 1st stage
Ethnicity	Categorical	- Bumiputera - Chinese - Indian and others (Indian)	Q. 1.3 in 1st stage
Age	Continuous/ Categorical	- 23 – 24 years old - 25 – 28 years old	Q. 1.1 in 1st stage
Marital status	Categorical	- Single - Married	Q. 30 in 3rd stage
Place in which an individual grew up	Categorical	- Rural - Small town - Large town	Q. 1.13 in 1st stage
Family background			
Parent's highest education attainment[4]	Categorical	- No formal schooling - Primary - Lower secondary - Upper secondary College/Polytechnic (College) - University	Q. 1.8 in 1st stage
Parent's occupational group[4]	Categorical	- Professional - Administrative - Clerical - Sales - Service - Agricultural - Production	Q. 1.9 in 1st stage
Parent's occupation ISEI[4]	Continuous/ Categorical	- 16 – 30 - 31 – 50 - 51 – 90	Q. 1.9 in 1st stage
Parent's monthly income[4]	Categorical	- No income - RM 500 and below - RM501 – RM1000 - RM1001 – RM1500 - RM1501 – RM2000 - RM2001 – RM3000 - RM3001 and above	Q. 1.11 in 1st stage

Family size (including the respondents)	Continuous/ Categorical	- 1 – 4 - 5 – 8 - 9 and above	Q. 1.5 in first stage
Birth order	Categorical	- Eldest - Not the eldest	Q. 1.6 in 1st stage
Human capital investment			
Formal education and training			
Years of academic education	Continuous/ Categorical	- 11 years[3] (Form 5 education) - 11.01 – 13 years (pre-university) - 13.01 – 16 years (college/uni.) - 16.01 years and above (Master/ PhD)	Aggregate years for all the courses in Q. 4 in 3rd stage
Highest qualification acquired	Categorical	- Secondary school - Certificate - Diploma - Professional or semi professional courses - Degree or higher	The highest level of qualification acquired in Q. 4 in 3rd stage
SRP aggregate	Continuous/ Categorical	- 5 – 10 aggregate - 11 – 15 aggregate - 16 – 20 aggregate - 21 – 40 aggregate	Q. 2.3 in 1st stage
SPM aggregate	Continuous/ Categorical	- 6 – 18 aggregate - 13 – 30 aggregate - 31 – 42 aggregate - 43 – 54 aggregate - Take MLVK	Q. 1e in 3rd stage
Pre-university result	Categorical	- Excellent - Above average - Average - Poor	Q. 3f in 3rd stage
Days of vocational training	Continuous/ Categorical	- No vocational training - 120 days or less - 121 – 240 days - 241 days or more	Aggregate days for all the courses in Q. 5 in 3rd stage
Training provided by employer and job tenure			
Days of training provided by employer	Continuous/ Categorical	- No training provided by employer - Have training provided by employer	Aggregate days for all the training in Q. 6 in 3rd stage
First job tenure	Continuous/ Categorical	- 6 months or less - 7 – 12 months - 13 – 24 months - More than 2 years	Q. 16 in 3rd stage

Employment			
First job occupational group	Categorical	- Professional - Administrative - Clerical - Sales - Service - Agricultural - Production	Q. 16 in 3rd stage
First job ISEI score	Continuous/ Categorical	- 16 – 40 - 41 – 60 - 61 – 90	Q. 16 in 3rd stage
First job starting monthly income	Continuous/ Categorical	- RM500 and less - RM501 – RM1000 - RM1001 – RM1500 - RM1501 and above	Q. 16 in 3rd stage
Years of working experience	Continuous/ Categorical	- 1 year or less - 2 – 3 years - 4 – 5 years - 6 years or above	Q. 17b in 3rd stage
Number of jobs held	Continuous/ Categorical	- 1 - 2 - 3 - 4 and above	Q. 17a in 3rd stage
Job changing pattern	Categorical (for income mobility only)	- Change occupation, same employer - Change occupation, change employer - Same occupation, change employer - Same occupation, same employer	Q. 16 in 3rd stage

[1] If an individual does not further studies after secondary school education, the first occupation is the occupation in the first row in Question 16. If he furthered studies, the first occupation is recorded in the second row.

[2] If an individual does not further studies after secondary school education, the first occupations income is given in the first row in Question 16. If he furthered studies, the first occupation's income is recorded in the second row.

[3] Although respondents, who finishes their primary school education in Chinese or Tamil National Type Primary Schools, have to go through one extra year of remove classes before they can proceed to Form 1 in secondary school, this extra year is not considered as one additional year of formal education. This year is mainly aimed to help to enhance the Bahasa Melayu competency of those coming from national school, as the main language used during the secondary school education system is Bahasa Melayu, which is different from the languages use during their primary school. So, every respondent have at least 11 years of formal education.

[4] For respondents whose guardian is both father and mother, father's information is used. For respondents with single parent, the single parent's information is used. If the respondents' present guardian is not their parent, then the guardian's information is taken.

3.6 Independent variables

The list of independent variables used is included in Table 3.2.

3.6.1 Demographic variables

The demographic variables of interest are gender, ethnicity, age, marital status and location in which an individual grew up.

3.6.2 Family background

The family background information that is relevant is parent's educational level, parent's occupation, parent's income, family size and birth order. For respondents whose guardian is both father and mother, father's education, occupation and income level are taken for evaluation. For respondents with single parent, the single parent's information is used. If the respondents' present guardian is not their parent, then the guardian's information is taken. For the parent's occupation, the original ISCO code is changed into the ISEI score to depict the socioeconomic status of the parent.

3.6.3 Human capital investment

3.6.3.1 Formal education and vocational training

Due to the incomparability of different education system and wide variety of education routes offered after the basic five years of secondary school education, the amount of education completed is used as a proxy to measure the formal education an individual receives (Ganzeboom et al., 1992).

The duration of formal education is adjusted for those who study part time. If the duration taken is longer for those who studied part time, compared to those who studied on a full time basis (on average), the duration recorded is the duration taken to finish the course on a full time basis, not the duration given by the respondents. However, if the duration is shorter or the same as on the full time basis, the duration considered is the length that is given by the respondents. If respondents took a course yet did not reveal any information on the duration of the courses, only those courses with the information of duration are considered. The duration is measured by year. It is divided into four categories, based on the different duration for every level of education on average.

In addition, the highest qualification acquired is another variable used to measure the amount of formal education an individual receives. Although all the qualifications acquired are recorded for different courses attended by respondents, only the highest qualification is used.

For academic performance, SRP aggregate is used to measure the performance for all respondents. SPM aggregate is another measurement for most of the respondents, except for those who sit for NITTCB. For SRP and SPM aggregates, lower aggregate means better result. For those who proceeded to pre-university, the pre-university result are also used as a barometer. Due to the difference of assessment systems in the pre-university programmes attended by the respondents, pre-university result is divided into four categories. They are excellent, above average, average and poor.

For vocational training, duration of training, measured in days, is used as a measurement in this study.

3.6.3.2 Training provided by employer and job tenure

Training provided by employer is measured using the duration of the training attended, which is measured in term of days. The job tenure is measured by tenure in the first job in days.

3.6.4 Employment

The relevant first job criteria are the occupational group, the ISEI score, starting monthly income. Other aspects of employment looked into are years of working experience, number of jobs held and job changing pattern.

3.6.5 Motivation and work value

Motivation can be measured using level of agreement or disagreement to various statements in Question 2.24 in first stage survey. It is analysed using principal component analysis. The relevant statements are shown in Table 3.3.

There are two work values that are relevant to this study. They are the work value in school and the work value in labour market. The work value in school is measured using level of importance of certain factors that provide job satisfaction, which were asked in Question 3.5 in the first stage questionnaire. The work value in the labour market is measured using level of importance of certain ideal job characteristics, which were asked in Question 14 in the third stage questionnaire. The relevant factors and characteristics are shown in Table 3.3.

Table 3.3 Statements that are relevant in analysing an individual's motivation, work value in school and labour market

Motivation	
1	I like to do well in examinations so that my teachers will be pleased.

2	Examination success is what I aim for throughout my school learning.
3	I study very hard in class to compete with my classmates.
4	I do not care very much about doing well in my studies.
5	Studying gives me a lot of personal satisfaction.
6	I work very hard in class to please my parents.
7	I like to do well in examination to improve my job qualification.
Work value in school	
1	Interesting work
2	Using special talents
3	Creative work
4	No supervision
5	Opportunities to further studies
6	Opportunity for improving competence
7	Opportunity to be helpful to the community
8	Work with people
9	Good income
10	Opportunity for travel
11	Supervise others
12	Secure future
13	Time for family/ friends
14	Using skills learned from schooling
15	Proximity to working place of spouse
Work value in labour market	
1	Have sufficient time for myself/ family
2	Get to perform challenging duties that give satisfaction
3	Job with minimum pressure or tension
4	Good working environment (example: good air circulation and lighting)
5	Have good working relationship with head
6	Secure occupation
7	Freedom to use own approach to perform job
8	Work with co-operative people
9	Head that involves subordinates to make decision
10	Have chance to contribute to the success of company/ organisation
11	Have opportunity to get better pay
12	Have opportunity to serve the country
13	Enable myself to stay in my dream area
14	Have opportunity to progress to higher position
15	Occupation with variety and adventurous element
16	Occupation in a prestigious and progressive company/ organisation
17	Have opportunity to help others
18	Have clear job requirements
19	Work in a big organisation

3.7 Computer programmes used

Most of the statistical analyses were done using the Statistical Package for Social Science (SPSS). All the graphical presentations are also produced using this package. Regression with White's heteroscedasticity-consistent variances and standard error was carried out using STATA. The analyses are based on weighted data.

3.8 Methodology used

Data analysis is divided into four main sections. They are exploratory data analysis, testing of mean difference, data reduction and explanatory data analysis. For all the statistical tests carried out, the chosen level of significance (α) is 0.10.

3.8.1 Exploratory data analysis

Explanatory data analysis is used to examine the pattern and the distribution of the data (Tukey, 1977). Exploratory data analysis focuses more on visual presentation of the data. (Cooper and Schindler, 1998). The graphical presentation method chosen is box-plot.

Box-plot

In studying the distribution for different categories in a factor, box-plot is chosen. Box-plot provides the five-number summary of a distribution. The five-number summary refers to first quartile (Q_1), median, third quartile (Q_3), minimum and maximum. Box-plot indicates clearly the range of the central 50% observation, which is called the interquartile range ($Q_3 - Q_1$). The whiskers (the vertical line extending on both side of the box) indicate the range of observations that fall within 1.5 times of the interquartile.

Median and quartiles are presented, rather than mean and standard deviation, as they are more 'resistant' and insensitive to the change and the abnormalities in a data set (Cooper and Schindler, 1998).

In addition, box-plot is helpful in telling the shape of the distribution. For a left-skewed distribution (negatively skewed), there is a long upper whisker and the median rests in the lower end of the box. Most of the observations concentrate in the higher end of the scale. For right-skewed distribution (positively skewed), there is a long lower whisker and the median lies in the upper side of the box. Most of the observations cluster in the lower end of the scale. If the length of the upper whisker and the lower whisker is almost equal and the median divides the box into half, this is a symmetrical distribution.

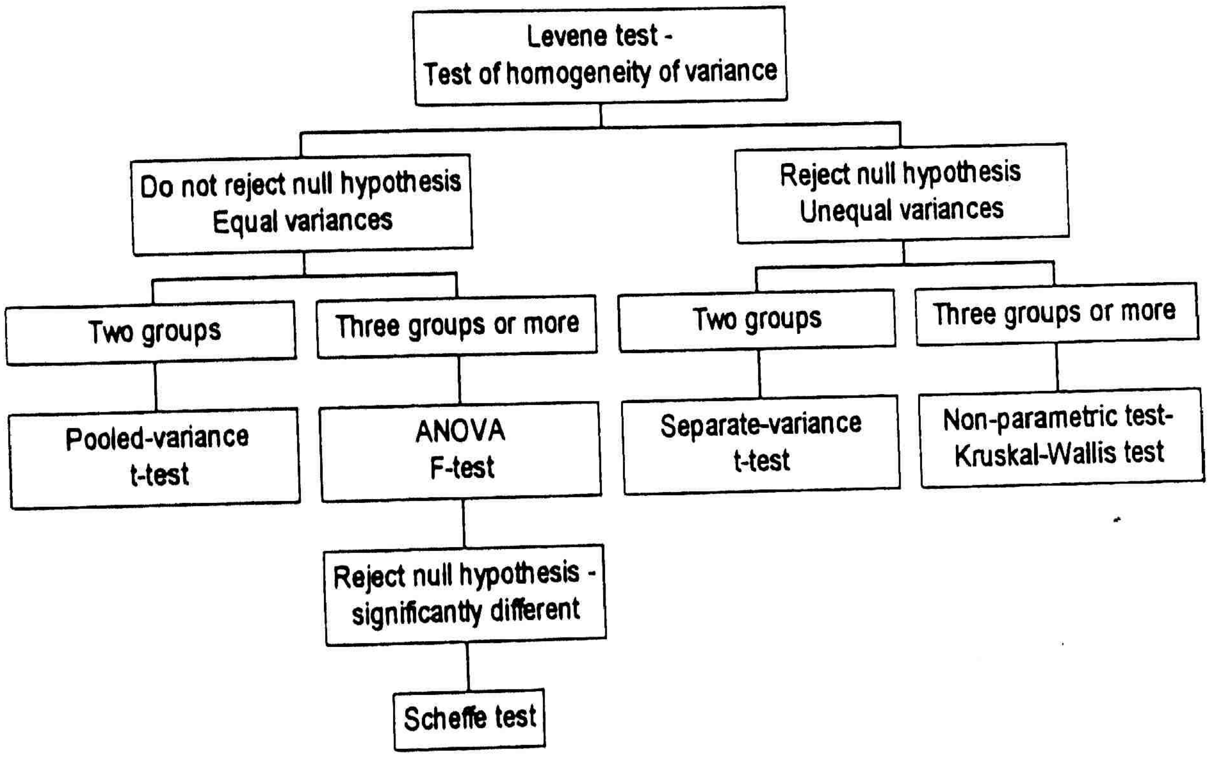
It is important to observe the shape of a distribution before making comparison among categories in a factor. If one of the distributions is left skewed and the other is right skewed, the difference in the measures of central tendency may not mean a lot. Nevertheless, if both are right skewed or left skewed, the difference may be a better indicator.

3.8.2 Testing of mean difference

To verify the differences observed in exploratory data analysis, the difference in mean is tested. A structure of procedures need to be carried out, as shown in Figure 3.1. Before ascertaining the mean difference of the aspect studied among categories of a factor, Levene test is carried out to test the homogeneity of variance.If the variances are significantly different, a non-paramatric test – Kruskal-Wallis test is conducted for a factor with three or more groups. If there are two categories, independent samples t-test is preferred for testing mean differences.

If the variances are homogeneous, ANOVA F-test is used to verify the differences in mean. If the difference is statistically significant, Scheffe test is applied to identify the pair that is significantly different. To confirm the average difference between the ISEI score or income for the first job and current job, paired samples t-test is used.

Figure 3.1 Structure of testing of mean difference



The symbols used in general are:

H_o = null hypothesis

H_a = alternative hypothesis

X_{ij} = j^{th} observation in group i

N = total sample size

n_i = sample size for group i

c = number of groups in a factor

Sp = pooled variance

MST = mean square treatment

$\overline{X_{ij}}$= sample mean for group i

μ_i = population mean for group i

σ_i^2= population variance for group i

X_i= sample mean for group i

S_i^2= sample variance for group i

MSE = mean square error

d = $X_1 - X_2$

T_i^2= square of the sum of ranks assigned to group i

df = degree of freedom

Levene test

Levene test is used to assess the homogeneity of variances among groups in a factor. The violation of homogeneity of variance affects the result of F-test and t-test seriously. Levene test is chosen as other tests (e.g. Bartlett's test and Hartley's test) are very sensitive to the violations of normality distribution and the ensuing ANOVA procedure is robust to the normality assumption (Jobson, 1991). So, Levene test is preferred.

$H_o : \sigma_1^2 = \sigma_2^2 = \ldots. = \sigma_c^2$

H_a : Not all σ_i^2 are equal

Test statistics

$$F = \frac{(\sum_{i=1}^{c} n_i(\overline{X}_i - \overline{X}_{ij})^2)/(c-1)}{(\sum_{i=1}^{c}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_i)^2)/(n-c)}$$

df for numerator = c – 1

df for denominator = n – c

If the null hypothesis is rejected, it may be concluded that the variances of groups in a factor differ significantly.

Assumption: All the group populations are normally distributed and the samples from each population are random and independent.

Independent samples t – test

To test the differences between the group population mean for a factor with two independent groups, a suitable independent samples t-test has to be chosen between the pooled-variance t-test and separate-variance t-test. For equal-variance groups, pooled variance t-test is more suitable. For unequal-variance categories, separate-variance t-test is more appropriate.

$H_o : \mu_1 = \mu_2$

$H_a : \mu_1 \neq \mu_2$

If the null hypothesis is rejected, it may be concluded that there is evidence of a mean difference in the aspect studied for two groups.

Assumptions: Random and independent samples from normal populations.

(i) Pooled-variance t-test

Test statistics

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$df = n_1 + n_2 - 2$

$$S_p = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

(ii) Separate variance t-test

Test statistics

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$df = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$$

One-way ANOVA F-test

Analysis of variance (ANOVA) is used to test the difference of group population mean among three or more groups in a factor.

$H_o : \mu_1 = \mu_2 \ldots\ldots = \mu_c$

H_a : Not all the μ_i are equal

Test statistics

$$F = \frac{MST}{MSE}$$

$$MST = \frac{\sum_{i=1}^{c} n_i(\overline{X}_i - \overline{X}_{ij})^2}{c - 1}$$

$$MSE = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_i)^2}{n - c}$$

df for numerator = c – 1

df for denominator = n – c

If the null hypothesis is rejected, it may be concluded that there is significant mean difference in the aspect studied for at least two groups in the factor.

Assumption: All the group populations are normally distributed and the samples from each population are random and independent.

Kruskal-Wallis test

If the groups' variances are significantly different, Kruskal-Wallis is the non-parametric alternative to compare the location of more than two groups. It is based on ranked data.

H_o : The population relative frequency distributions for all c groups are identical

H_a : At least two populations relative frequency differ.

Test statistics

$$H = \left(\frac{12}{n(n+1)}\sum_{i=1}^{c}\frac{T_i^2}{n_i} \right) - 3(n+1)$$

Reject null hypothesis if $H > \chi^2_\alpha$ with (c – 1) df.

If the null hypothesis is rejected, it may be concluded that there is sufficient evidence to say that the aspect studied differs among the groups.

Assumption: All the samples from each group population are random and independent.

Multiple comparisons - Scheffe test

Scheffe test is useful in identifying the pair that is significantly different after the null hypothesis in F-test is rejected. Scheffe test verifies all the possible pairs of mean difference. It is the most conservative method to ascertain the differences between a pair of means (Jobson, 1991). For Scheffe test, the sample sizes may not be all equal. In addition, if the assumptions of normality and equal population variances are not satisfied, this test may still be helpful.

$H_o : \mu_a = \mu_b$

$H_a : \mu_a \neq \mu_b$

Reject null hypothesis if

$$\left|\overline{X}_a - \overline{X}_b\right| > S\sqrt{MSE\left(\frac{1}{n_a} + \frac{1}{n_b}\right)}$$

$$S = \sqrt{(c-1)F_{c-1,n-c;\alpha}}$$

If the null hypothesis is rejected, it may be concluded that the mean of group a is significantly different from group b.

Paired samples t-test

This test is used to confirm the difference of two dependent samples. For example, the difference of ISEI scores and income for the first and current job. It is only fair to compare an individual's first job ISEI score and income with the same individual's current job ISEI score and income.

$H_o : \mu_d = 0$

$H_a : \mu_d \neq 0$

Test statistics

$$t = \frac{\bar{d}}{\frac{S_d}{\sqrt{n}}}$$

df = n – 1 and $d = X_1 - X_2$

If the null hypothesis is rejected, it may be concluded that the mean difference in the aspect studied for first and current job is statistically significant.

Assumptions: Population paired difference is normally distributed.

3.8.3 Data reduction

The data reduction method used is principal component analysis. It is aimed to reduce the large number of independent variable used to measure a subjective idea. A new and smaller set of variables is created using principal component analysis. For detail, please refer to Hair et al., 1998 and Johnson, 1998.

Principal component analysis is concerned more with explaining variability in the related variables and produces an orthogonal transformation of the variables, which does not depend on any underlying models.

Principal component analysis is applied to the correlation matrix of the variables. This is equivalent to applying the procedure on standardised data, rather than raw data. This means that it is assumed that all the variables are equally important. This assumption is reasonable as all the variables included in the principal component analysis explain different aspect of a subjective idea.

The new variables that are formed is called principal components. These components are not correlated and the first principal component accounts for as much as possible the variability in the variables. It accounts for the most variability, compared to the succeeding components.

The procedure for principal component analysis is as followed:

- Checking the correlation and sampling adequacy

  The variable corelation matrix is checked to see whether there is a substantial number of correlations that is larger than 0.3 (Hair et al., 1998). If there is no substantial number of correlation larger than 0.3, this may indicate that principal component analysis is inappropriate. This can be further checked using the determinant of the correlation matrix. A determinant close to zero indicates that linear dependencies is found among the related variables and principal component analysis is suitable (Johnson, 1998).

  The degree of inter-correlation among variables can be measured using individual measure of sampling adequacy for each variable. The closer this value is to 1, the better this variable is being explained by other variables. So, a large value of measure of sampling adequacy indicates that this variable is suitable in this analysis. The threshold level suggested is 0.5. In addition, Kaiser-Meyer-Olkin measure of sampling adequacy is an overall measure of sampling adequacy. Any value below 0.5 indicates that principal component analysis is unsuitable.

Another statistical test is Bartlett test of sphericity. This test is used with the assumption that the data is from a normal multivariate distribution. The null hypothesis is that the population correlation matrix is an identity matrix, which means all the variables are uncorrelated.

- After justifying that principal component analysis, the number of components to be extracted from the variable need to be determined. Since the variance of each standardised variable is 1 when the principal component is applied on correlation matrix, any component that cannot account for more variance than a single variable is dropped. So, the number of components to be extracted is the number of eigenvalue that are larger than 1.
- A new set of principal component scores are computed. These components are then included in the multiple regression model as independent variables. Only components that account for more than 10% of the variance of the variables considered in the principal component analysis are included in the regression model.

  In most cases, the principal component scores generated are not meaningful. So, it is expected that the scores are not interpretable to have any meaningful implication (Johnson, 1998).

3.8.4 Explanatory data analysis

The main explanatory data analysis used is multiple regression, which is aimed to identify the attributes that affect the career advancement an individual achieves through occupational and income mobility. In addition, it is possible to differentiate the importance and the net effect of the attributes. Since career advancement is estimated ing occupational mobility and income mobility, two multiple regression models are timated.

Adjusted multiple correlation coefficient (R square) and Standard Error of gression (SER) are proxies for the goodness of fit of the model. A high adjusted R juare and a low SER indicate a good fit. As to measuring the relative importance of the idependent variables in a model, beta coefficient is used. The higher the absolute value, ie more important is the variable, compared to other variables in the model.

ssumptions of classical multiple regression model

he multiple regression model is

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

ssumptions of this model are as below (refer to Gujarati, 1995):

- Zero mean value of ε_i –

$$E\left(\varepsilon_i \middle| X_1 \ldots X_p\right) = 0$$

- No serial correlation – cov $(\varepsilon_i, \varepsilon_j) = 0, i \neq j$
- Homoscedasticity (equal variance) – $var(\varepsilon_i) = \sigma^2$
- Zero covariance between ε_i and each X variable - cov (ε_i, X_{2i}) = cov (ε_i, X_{3i}) = ... = cov $(\varepsilon_i, X_{pi}) = 0$
- The model is correctly specified, implying no specification error
- No exact collinearity between the independent variables, Xs – low multicollinearity

Building the regression model

This section explains the stages to build the final model for occupational mobility score and income mobility score. The variables to be included in a model should not be highly correlated to avoid the problem of multicollinearity.

Firstly, a regression model consisting the demographic characteristics, human capital investment, employment and family background is built. If there is more than one variable measuring a certain aspect, one of the variables is selected. As to the selection of variable, one of the variables is put into the model at a time. The selected variable is statistically significant and has an interpretable sign and the model built has the highest adjusted R square or lowest SER. Secondly, the principal component scores for motivation as well as work value in school and work value in labour market are included into the model. Thirdly, the interaction terms are included.

A categorical variable is converted into a dummy variable with two groups, with one group taking the value 1 and the other (base) group taking the value 0. The identification of the two groups for each variable was based on the results in the univariate analysis. The basic rule was to identify the groups in such a way so as to capture the greatest difference in career advancement.

Checking the assumptions of classical multiple regression model

a) Multicollinearity

Multicollinearity refers to the correlation among independent variables. It is a data problem, and not a model problem. Multicollinearity always exists, however it is harmful to the result if the collinearity is very high. It makes precise estimation difficult. The contribution of each independent variable is compounded. Hence, it is tough to give individual explanation for the effect of each independent variable on career advancement through mobility (refer to Hair et al., 1998).

Detection of multicollinearity (refer to Gujarati, 1995)

- Model with high R square and highly significant F test in most of the cases but very few individual t-tests show that the slope coefficients are significant.
- High correlation among independent variables. This condition is a sufficient criteria for the existence of multicollinearity but it is not a necessary condition because it may exist even when there is low correlation. This measure may not be able to detect multicollinearity if three or more independent variables are correlated while there is no high relationship between any two variables (Belsley et al., 1980).
- A priori research shows that there is high correlation between these variables.
- Variance inflation factor (VIF) is a measure to assess pairwise or multiple collinearity. High VIF (> 10.0) implies serious multicollinearity. A high VIF indicates that R_j^2 near to unity, indicating that most of the information provided by this variable can be explained using other variables in the model.

$$VIF = \frac{1}{1 - R_j^2}$$

where R_j^2 is the multiple correlation coefficient of X_i regressed on other independent variables.

b) Normality of the error term

Non-normality in the error term is most frequently encountered. Fortunately, moderate departure from the normality assumption has limited effect on the result of the statistical tests and the prediction. Hence, regression is robust to the violation of the normality assumption. Nevertheless, it is important to check the degree of violation as serious violation may incur problem to the regression model.

Histogram and normal probability plot are chosen to assess this assumption. If histogram resembles the normal curve imposed, this indicates the normality assumption is not violated. The normal probability plot compares the distribution of the residual with a normal distribution. The normal distribution forms a straight diagonal line. Then, the residual is compared with this diagonal line. If the residuals closely scatter around the straight line, the distribution is normal (refer Hair et al., 1998).

c) Heteroscedasticity

Heteroscedasticity refers to the problem of unequal variance of the residual for all observations. If normal testing procedure is carried out with the existence of heteroscedasticity, the result is misleading (Gujarati, 1995).

Detecting heteroscedasticity (Refer to Gujarati, 1995)

- Plot the residual squared, ε_i^2, against the estimated Y_i from the multiple regression model. If a pattern is noted, this indicates the problem of heteroscedasticity.

- White's general heteroscedasticity test is a formal measure to detect the violation of homoscedasticity. This test is chosen as it is easy to implement and does not depend on the normality assumption.

  The null hypothesis states that there is no heteroscedasticity while the alternative hypothesis assumes some unknown general form of heteroscedasticity. The test statistics is computed using an auxiliary regression – the squared residual on all possible cross products of the independent variable. For example,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

  The auxiliary regression is

$$\hat{\varepsilon}_i^2 = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 X_{2i}^2 + \alpha_5 X_{3i}^2 + \alpha_6 X_{2i} X_{3i} + v_i$$

  Reject the null hypothesis if (n.R square) > χ^2 with degree of freedom taking the value of number of independent variables in the auxiliary regression (except intercept). The R square is obtained from the auxiliary regression.

Diagnostics

The purpose of this section is to identify the observation that is considered as outlier or influential data. After the outliers and influential observations are identified, their distributions are studied. Then, they are removed from the model temporarily and the model is being estimated again without those observations. The model without those observations is then compared with the original model. If a huge difference is noted and the characteristics of those observations differ from others, those observations are removed permanently and the re-estimated model is accepted.

Outlier

Outliers are observations that has unique characteristics and distinct from other observations. Since regression model involving more than two variables is used in this study, Mahalanobis D^2 measure is used to examine each observation across a set of variables. It measures the distance in multidimensional space of each observation in the data set from the mean centre of the observations. The rule to determine outliers: a small number of observations with the highest Mahalanobis D^2 that are two or three times to the next highest value.

Influential observation

Influential observations are observations that have disproportionate influence on the dependent variable in a regression model. It is not a must that an influential observation must be an outlier, although most of the time outliers are influential observations too. In this study, Cook's distance is chosen to detect influential observations (refer to Hair et al., 1998).

Cook's distance (D_i) captures the impact of the size of change of the dependent variable if the observation is deleted and the distance of the observation from other observations. The threshold level is $4/(n - k - 1)$, where k is the number of independent variables excluding intercept and n is the sample size.