



University
of Glasgow

Letham, Collette Alexis (2012) Documenting & imputing missing values in a longitudinal survey of students' personal attributes. MSc(R) thesis

<http://theses.gla.ac.uk/4545/>

Copyright and moral rights for this thesis are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the Author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the Author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.



University
of Glasgow

**Documenting & Imputing Missing Values in a
Longitudinal Survey of Students' Personal Attributes**

Collette Alexis Letham

*A Dissertation Submitted to the
University of Glasgow
for the degree of
Master of Science in Statistics*

School of Mathematics and Statistics

September 2012

© Collette Alexis Letham

Abstract

The University of Glasgow is currently engaged in a programme of action designed to reduce the proportion of students who withdraw from the university during their first year. Student retention is a cause for concern for higher education institutions in terms of reputation and funding.

Previously, researchers have suggested that early withdrawal from university is linked to personal attributes. A questionnaire to explore this was designed consisting of 5 standard psychometric scales measuring respectively mindset, self efficacy, self esteem, resilience and hope. All new entrants to the University of Glasgow in September/October 2009 were invited to take part in a study of these personal attributes. 1,098 (20%) new undergraduates and 407 (10%) new postgraduates agreed, and filled in the questionnaire while pre-registering on the university's computerized registration system (WebSURF). At random, half of the students who took part at baseline were invited to complete the same survey again at the end of teaching in Semester 1 and the other half at the end of teaching in Semester 2.

The results obtained on the psychometric scales were linked to routinely-collected data about the same students' background and their continuation and progression at the end of first year. The aim was to investigate the influence of personal attributes, either on their own or in conjunction with demographic variables, on the continuation and progression of students.

A common problem encountered in this study is that data were missing. It is important that the reasons why data are missing are taken into account and that missing data is dealt with, as far as possible, in a way that does not lead to biased results and invalid inferences. For this reason, it was decided not to rely on the results of a complete case analysis, but to use multiple imputation to fill in the missing values and then repeat the analysis using the completed datasets as well.

Chapter 2 provides a review of the psychometric scales used in this study. The characteristics of missing data and methods to handle missing data are described. Also in Chapter 2, the theory of various statistical methods used in this analysis is illustrated in detail.

In Chapter 3 the completeness of the questionnaire dataset is documented by examining the rates of non-response. The completeness of the questionnaire is also examined to establish if any of the demographic variables such as Sex, Age, Domicile, Faculty and Socio-Economic Class are associated with it. A higher proportion of older than younger undergraduate students completed the questionnaire fully, and more students in a professional faculty than students in a non-professional faculty completed it.

The complete case analysis to explore the effect of demographic variables and personal attributes on the outcome of first year for undergraduate students is detailed in Chapter 4. For whether or not first year students continued at the University of Glasgow after first year neither the baseline personal attribute scores nor the difference in personal attribute scores were found to be statistically significant. The change in self esteem score in the course of first year was seen to be a significant predictor of whether or not first year students progressed at the University of Glasgow after first year.

Chapter 5 focuses on various ways in which that imputation was applied to fill in missing values of the baseline personal attribute scores and the difference in personal attribute scores. However, even after imputing the personal attribute data, neither the baseline personal attribute scores nor the difference in personal attribute scores were found to be statistically significant predictors of Continuation or Progression.

Chapter 6 includes a summary of the results of this thesis and discusses the limitations and further work that could be implemented.

Acknowledgements

The first and most important thank you must go to my supervisor Prof. John H. McColl. Firstly for giving me the opportunity to do an MSc but also for your guidance, patience, invaluable expertise and supervision, without it this thesis would not be possible. It has been both a pleasure and a privilege to work with him.

I would also like to extend my thanks to Alison Browitt whose advice and study materials have been invaluable throughout.

I would like to acknowledge the University Of Glasgow for providing the data for this project and the Engineering and Physical Sciences Research Council (EPSRC) for funding me throughout this research.

Thanks are also due to my fellow Msc'ers Kathryn, Stephen, Mhairi, Greg, Laura and Andisheh for providing well needed laughter, games, shenanigans and most importantly the endless cups of tea & biscuits!

I would also like to thank my friends and family for providing me with support and encouragement, whether it was for the distractions when needed or someone to vent to, even when they didn't quite understand, it has been very much appreciated.

Finally, I would like to thank my parents for their belief, encouragement and reminding me to never give up. I would never have gotten this far without your support.

Contents

Abstract.....	i
Acknowledgements	iii
Contents	iv
List of Tables	vi
List of Figures.....	x
Chapter 1 Introduction.....	1
1.1 Introduction.....	1
1.2 Aims.....	3
Chapter 2 Methods	4
2.1 Dataset.....	4
2.1.1 Personal Attribute Scales	4
2.1.1.1 Mindset	5
2.1.1.2 Self Efficacy.....	6
2.1.1.3 Self Esteem	7
2.1.1.4 Resilience.....	7
2.1.1.5 Hope.....	8
2.1.2 Data Variables.....	10
2.2 Missing Data	11
2.2.1 Missing Data Mechanisms	11
2.2.2 Method of Handling Missing Data	13
2.2.2.1 Complete Case Analysis	13
2.2.2.2 Imputation	14
2.2.2.3 EM Algorithm.....	17
2.3 Methods of Analysis	17
2.3.1 Fisher’s Exact Test.....	17
2.3.2 Binary Logistic Regression.....	18
2.3.3 Model Building	19
2.3.4 Hosmer-Lemeshow Test	21
2.4 Statistical Programs	22

Chapter 3 Data Description	23
3.1 Exclusion Criteria	23
3.2 Descriptive Statistics.....	25
3.3 Non Response	28
Chapter 4 Complete Case Analysis	45
4.1 Continuation at Baseline	45
4.2 Progression at Baseline	51
4.3 Continuation at Baseline and Semester 1/Semester 2 for Difference in Personal Attribute Scores	56
4.4 Progression at Baseline and Semester 1/ Semester 2 for Difference in Personal Attribute Scores	60
Chapter 5 Multiple Imputation	67
5.1 Imputing Scale Level Data at Baseline	68
5.2 Imputing Item Level Data at Baseline	71
5.2.1 Continuation for Imputed Item Level Data at Baseline	71
5.2.2 Progression for Imputed Item level Data at Baseline	75
5.3 Imputing at Baseline, Semester 1 and Semester 2	79
5.3.1 Continuation for Imputed Data at Baseline and Semester 1/ Semester 2 for Difference in Personal Attribute Scores.....	81
5.3.2 Progression for Imputed Data at Baseline and Semester 1/ Semester 2 for Difference in Personal Attribute Scores.....	85
Chapter 6 Discussion	90
6.1 Conclusions.....	90
6.2 Limitations of the Study & Further Work.....	94
Bibliography	96
Appendix A Questionnaire.....	98
A.1 Personal Attributes Questionnaire	98
Appendix B Programming Code	105
B.1 Example Model Building Code.....	105
B.2 Code for Imputing Scale level Data at Baseline	110
B.3 Code for Imputing Item level Data at Baseline.....	111
B.4 Code for Imputing at Baseline and Semester – Step 1	112
B.5 Code for Imputing at Baseline and Semester – Step 2	115

List of Tables

Table 2.1: Theories of Intelligence Scale	5
Table 2.2: The General Self Efficacy Scale	6
Table 2.3: Rosenberg's Self Esteem Scale	7
Table 2.4: Ego Resiliency Scale	8
Table 2.5: Trait Hope Scale	9
Table 3.1 Table of Excluded Data	25
Table 3.2: Demographic Variables by Time Point for Undergraduates	26
Table 3.3: Personal Attributes by Time Point for Undergraduates.....	26
Table 3.4: Demographic Variables by Time Point for Postgraduates	27
Table 3.5: Demographic Variables by Time Point for Postgraduates	27
Table 3.6: Rate of Wave Non-Response.....	28
Table 3.7: Rate of Item Non-Response & Personal Attribute Scale Non-Response	30
Table 3.8: Fisher's Exact Test of Completed Personal Attribute Scales by Sex at Baseline for Undergraduates	32
Table 3.9: Fisher's Exact Test of Completed Personal Attribute Scales by Sex at Baseline for Postgraduates	33
Table 3.10: Fisher's Exact Test of Completed Questionnaires by Demographic variables at Baseline for Undergraduates.....	35
Table 3.11: Fisher's Exact Test of Completed Questionnaires by Demographic variables at Baseline for Postgraduates.....	36

Table 3.12: Univariate Logistic Regression of Completion by Demographic Variables at Baseline for Undergraduates.....	38
Table 3.13: Models for Completion at Baseline for Undergraduates	39
Table 3.14: Logistic Regression of Completion for Age & Faculty at Baseline for Undergraduates	41
Table 3.15: Univariate Logistic Regression of Completion by Demographic Variables at Baseline for Postgraduates.....	42
Table 3.16: Models for Completion at Baseline for Postgraduates	43
Table 4.1: Continuation by Demographic Variables at Baseline.....	46
Table 4.2: Continuation by Personal Attributes at Baseline	47
Table 4.3: Some Models for Continuation at Baseline	49
Table 4.4: Logistic Regression of Continuation for Faculty and for Sex & Faculty & Age at Baseline.....	50
Table 4.5: Progression by Demographic Variables by at Baseline.....	52
Table 4.6: Progression by Personal Attributes at Baseline	53
Table 4.7: Some Models for Progression at Baseline	54
Table 4.8: Logistic Regression of Progression for Sex & Age at Baseline	55
Table 4.9: Continuation by Difference in Personal Attributes	57
Table 4.10: Some Models for Continuation at Baseline and Semester1/2	59
Table 4.11: Logistic Regression of Continuation for Sex & Difference in Self Esteem at Baseline and Semester1/2	60
Table 4.12: Difference in Personal Attributes by Progression	61
Table 4.13: Some Models for Progression at Baseline and Semester1/2	63

Table 4.14: Logistic Regression of Progression for Difference in Self Esteem at Baseline and Semester 1/2	64
Table 5.1: Plausibility for Scale Level Data at Baseline	69
Table 5.2: Example of Consistency	69
Table 5.3: Consistency for Scale Level Data at Baseline	70
Table 5.4: Direction of Non Consistent Values for Scale Level Data at Baseline	70
Table 5.5: Univariate Logistic Regression of Continuation by Personal Attribute Scales at Baseline.....	73
Table 5.6: Combined Estimates and Sampling Variability of Continuation by Personal Attribute Scales at Baseline	73
Table 5.7: Logistic Regression of Continuation for Sex & Faculty at Baseline.....	74
Table 5.8: Univariate Logistic Regression of Progression by Personal Attribute Scales at Baseline.....	77
Table 5.9: Combined Estimates and Sampling Variability by Personal Attribute Scales at Baseline.....	77
Table 5.10: Logistic Regression of Progression for Sex and for Sex & Age & Faculty at Baseline.....	78
Table 5.11: Univariate Logistic Regression of Continuation by Difference in Personal Attribute Scales at Baseline and Semester 1/2.....	83
Table 5.12: Combined Estimates and Sampling Variability by Difference in Personal Attribute Scales at Baseline and Semester 1/2.....	83
Table 5.13 : Results of Model Building for Continuation	84
Table 5.14: Univariate Logistic Regression of Progression by Difference in Personal Attribute Scales at Baseline and Semester 1/2.....	87

Table 5.15: Combined Estimates and Sampling Variability by Difference in Personal Attribute Scales at Baseline and Semester 1/2.....	87
Table 5.16: Results of Model building for Progression.....	88

List of Figures

Figure 4.1: Probability of Progression by Difference in Self Esteem Score	65
---	----

Chapter 1

Introduction

1.1 Introduction

A longitudinal study is defined as a study where experimental units (e.g. people or animals) are repeatedly measured over time (Diggle et. al 2002). Several variables of interest can be measured for each experimental unit at specific time points throughout the study. Missing data commonly occur in longitudinal studies, this is the case when one or more of the repeated measurements on an experimental unit within the study are incomplete. Careful analysis is required when data are missing in a longitudinal study, otherwise a bias can be introduced leading to misleading inferences. This thesis investigates the consequences of missing data for the analysis of a longitudinal study of student retention, recently conducted by researchers at the University of Glasgow.

For institutions of higher education, student retention has become a cause for concern in terms of reputation and funding. The loss of revenue for a higher education institute through unrealised tuition fees and alumni contributions is in the thousands for each student that withdraws (DeBerard et al, 2004). During 2008/2009 10.7% of students did not continue at the same higher education institute within the UK. Within Scotland this increased to 11.4% of students (Higher Education Statistics Agency). Due to this, the University of Glasgow is currently engaged in a programme of action to reduce the proportion of students who withdraw from the university during their first year. Research elsewhere (Bean & Eaton 2000) has suggested that early withdrawal may be linked to students' personal attributes and changes in these attributes during first year. To investigate the relationship between personal

attributes and first year outcome, a questionnaire was designed consisting of the 50 questions that make up five standard psychometric scales measuring respectively mindset, self-efficacy, self-esteem, resilience and hope.

All new entrants to the University of Glasgow in September/October 2009 were invited to take part in a study of personal attributes. 1,098 (20%) new undergraduates and 408 (10%) new postgraduates agreed, and filled in the questionnaire while pre-registering on the university's computerised registration system (WebSURF). The students who agreed to take part in the study at baseline were then invited to complete the same questionnaire again. Using stratified random sampling, half of them were invited to do this at the end of teaching in Semester 1 and the second half were invited at the end of teaching in Semester 2. Students were only asked to fill in one follow up questionnaire as it would be likely that students would be able to remember their responses if they were asked to repeat the questionnaire again. 220 undergraduates and 93 postgraduates agreed in Semester 1 and 165 undergraduates and 78 postgraduates agreed in Semester 2.

Each student's demographic details were also collected from the University's central database using their registration number: information on Faculty, Gender, Age, Domicile, and Attendance Status were collected from all students. In addition to these, Socio-Economic Class (SEC) and Qualifications on Entry were also collected on undergraduate students only. In November 2010, following the re-sit examination diet in August 2010, final first year results were added the University's central database allowing for information about continuation and progression for each student to be accessed. Ethical approval was granted by the Faculty of Information and Mathematical Sciences Ethics Committee in June 2009, including use of students' data.

A variety of missing values have occurred in this study including the following:

- Occasional questions being missed out by subjects while other questions had been answered within the same psychometric scale, meaning that the score for that psychometric scale could not be calculated. When a question has not been answered or missed out this is called Item non-response.
- Students missing out whole scales although they had completed other scales.
- Many of the students who completed the survey at baseline did not take part in the follow up questionnaire at Semester 1/Semester 2. This is known as Wave non-response.

In any study it is important that missing data is dealt with, as far as possible, in a way that does not lead to biased results and invalid inferences. It is also important to take into account why the data is missing and if the missingness is related to why the data is being analysed.

Three terms were first introduced by Rubin in 1976 for the different mechanisms that lead to missing data and whether or not missingness is associated with the underlying values in the dataset (Little and Rubin, 2002). The three types of missing data mechanisms are: Missing Completely at Random (MCAR), Missing at Random (MAR) and Not Missing at Random (NMAR). MCAR means that missingness does not depend on the missing or observed data, MAR means that missingness depends on the observed data but not the missing data and NMAR means that missingness depends on the missing data. Depending on which missing data mechanism is in operation the appropriate way to analyse the data is different.

1.2 Aims

This thesis aims to document the completeness of the questionnaire. The number of students who did and did not complete each item in each scale at each time point will be documented clearly identifying where Item non-response and Wave non-response occurs. The missingness will then be investigated to establish if it is related to any demographic variables such as sex, age, domicile, faculty and SEC.

The general literature on psychometric testing and the literature specific to the scales used in this study will be looked into to clarify how other researchers have dealt with missing items within otherwise completed scales. Multiple imputation will be applied as a structured alternative to these ad-hoc procedures.

The purpose of the study that produced these data was to investigate the influence of personal attributes on continuation and progression of students after the end of first year for Undergraduates. Therefore this thesis will compare the results obtained by analysing these data using complete cases only and using imputed datasets.

Chapter 2

Methods

2.1 Dataset

As described in section 1.1 students' personal attribute scale responses, demographic details, Socio-Economic Class, continuation and progression information were collected from all of the students who agreed to take part in the study. Section 2.1.1 describes what each personal attribute scale measures and how the score is calculated. Section 2.1.2 describes how the demographic details, SEC, continuation and progression were grouped and coded for this thesis.

2.1.1 Personal Attribute Scales

The 5 standard psychometric scales chosen to investigate the relationship between personal attributes and first year outcome respectively measure mindset, self-efficacy, self-esteem, resilience and hope. All of the psychometric scales chosen for the questionnaire were recommended by the Centre for Confidence & Well-being and are commonly used in an academic situation. These were determined out with the scope of this thesis. The 5 standard psychometric scales have had their reliability and validity investigated in numerous studies and are proven to have high reliability and validity. They are acceptable for use on the adult population and in longitudinal studies. For all of the psychometric scales, all items in the psychometric scale have to be answered for the score to be calculated.

2.1.1.1 Mindset

To measure mindset, ‘Theories of Intelligence Scale’ (Dweck, C.S, C. Chui, & Y. Hong, 1995) was used. The scale measures a person’s belief about their own abilities: their mindset. The first belief that is measured is that ability and intelligence is fixed and doesn’t change, this is a fixed mindset. The second believe is that ability is not a fixed entity and can grow and improve over time, this is a growth mindset.

- | |
|--|
| <ol style="list-style-type: none">1. You have a certain amount of intelligence, and you can’t really do much to change it.2. Your intelligence is something about you that you can’t change very much.3. To be honest, you can’t really change how intelligent you are.4. You can learn new things, but you can’t really change your basic intelligence |
|--|

Table 2.1: Theories of Intelligence Scale

Each item, shown in Table 2.1, is scored on a 6 point scale with responses: ‘strongly agree’ (1), ‘agree’ (2), ‘partially agree’ (3), ‘partially disagree’ (4), ‘disagree’ (5) and ‘strongly disagree’ (6). The score is calculated by taking the mean of the 4 scores, giving a score range of between 1 and 6. The developers of the scale take a score of 3 or below to be related to a fixed mindset and a score of 4 or above to be related to a growth mindset.

The Centre for Confidence and Well-being through their own work were concerned that the wording of these items could be misleading and be misinterpreted as a general statement about other people and not a person’s own mindset. Carol Dweck, one of the original authors of the scale, was consulted about the changes by the Centre for Confidence and Well-being and in her opinion said that they wouldn’t affect the reliability or validity.

Some rewording was made to the 4 items:

Question 1 was changed to *‘I have certain inbuilt talents, like sport or music, and I can’t do much to change what those talents are.’*

Question 2 was changed to *‘There are subjects, like maths or languages that I’m naturally good at, but others that I’m naturally poor at and I don’t think I could ever be good in.’*

Question 3 was changed to *‘To be honest, I don’t think I can change how intelligent I am.’*

Question 4 was changed to *‘Although I can learn new things, I can’t really change what my talents and abilities are.’*

2.1.1.2 Self Efficacy

Self efficacy was measured using ‘The General Self Efficacy Scale’ (Schwarzer, R. & M. Jerusalem, 1995) shown in Table 2.2. The scale does not measure a person’s level of self-efficacy in a specific area instead it measures a general belief. It measures a person’s belief that they can successfully perform an action required to reach their goals. It is a belief that they can learn or perform a novel or difficult task, or cope with adversity, in a variety of different situations.

- | |
|---|
| <ol style="list-style-type: none">1. I can always manage to solve difficult problems if I try hard enough.2. If someone opposes me, I can find the means and ways to get what I want.3. It is easy for me to stick to my aims and accomplish my goals.4. I am confident that I could deal efficiently with unexpected events.5. Thanks to my resourcefulness, I know how to handle unforeseen situations.6. I can solve most problems if I invest the necessary effort.7. I can remain calm when facing difficulties because I can rely on my coping abilities.8. When I am confronted with a problem, I can usually find several solutions.9. If I am in trouble, I can usually think of a solution.10. I can usually handle whatever comes my way. |
|---|

Table 2.2: The General Self Efficacy Scale

This is a 10 item scale where each item is scored on a 4 point scale with responses: ‘not true at all’ (1), ‘hardly true’ (2), ‘moderately true’ (3) and ‘exactly true’ (4). The score is calculated by taking the sum of the 10 scores, giving a score range of between 10 and 40. The higher the score the more efficacious the person perceives himself or herself to be.

2.1.1.3 Self Esteem

‘Rosenberg’s Self Esteem Scale’ (Rosenberg, M., 1965) was used to measure self esteem. The scale measures self esteem which is defined as a positive or negative orientation towards oneself. It is an overall evaluation of one's worth or value.

- | |
|--|
| <ol style="list-style-type: none">1. On the whole, I am satisfied with myself.2.* At times I think I am no good at all.3. I feel that I have a number of good qualities.4. I am able to do things as well as most other people.5.* I feel I do not have much to be proud of.6.* I certainly feel useless at times.7. I feel that I am a person of worth, at least equal with others.8.* I wish I could have more respect for myself.9.* All in all, I am inclined to feel that I am a failure.10. I take a positive attitude toward myself. |
|--|

* Negative Items

Table 2.3: Rosenberg’s Self Esteem Scale

This is a 10 item scale where each negative item (marked with a *) is scored on a 4 point scale with responses: ‘strongly agree’ (1), ‘agree’ (2), ‘disagree’ (3), and ‘strongly disagree’ (4) and each positive item is scored on a 4 point scale with responses: ‘strongly agree’ (4), ‘slightly agree’ (3), ‘slightly disagree’ (2), and ‘strongly disagree’ (1). The score is calculated by taking the sum of the 10 scores, giving a score range of between 10 and 40. The higher the score relates to the more self esteem the person has.

2.1.1.4 Resilience

To measure resilience, ‘Ego Resiliency Scale’ (Block J. & A. M. Karmen 1996) was used. The scale measures a person’s abilities to adapt flexibly to stressful or challenging events in life. It also measures the ability to endure and recover from difficult situations.

- | | |
|-----|---|
| 1. | I am generous with my friends. |
| 2. | I quickly get over and recover from being startled. |
| 3. | I enjoy dealing with new and unusual situations. |
| 4. | I usually succeed in making a favorable impression on people. |
| 5. | I enjoy trying new foods I have never tasted before. |
| 6. | I am regarded as a very energetic person. |
| 7. | I like different paths to familiar places. |
| 8. | I am more curious than most people. |
| 9. | Most of the people I meet are likeable. |
| 10. | I usually think carefully about something before acting. |
| 11. | I like to do new and different things. |
| 12. | My daily life is full of things that keep me interested. |
| 13. | I would be willing to describe myself as a pretty 'strong' personality. |
| 14. | I get over my anger at someone reasonably quick. |

Table 2.4: Ego Resiliency Scale

This is a 14 item scale, shown in Table 2.4, where each item is scored on a 4 point scale with responses: 'disagree strongly' (1), 'disagree' (2), 'agree' (3) and 'strongly agree' (4). The score is calculated by taking the mean of the 14 scores giving a score range of between 1 and 4. The higher the score relates to the more resilient the person is.

2.1.1.5 Hope

'Trait Hope Scale' (Snyder et al 1991) was used to measure Hope. The scale assesses a person's global level of hope and how they generally perceive themselves in goal pursuits across situational contexts. Hope is defined as "the process of thinking about one's goals along with the motivation to move toward those goals (agency) and the ways to achieve those goals (pathways)" (Snyder et al, 2002). Therefore this scale has two subscales, hope agency and hope pathway.

Hope agency relates to a person's perception of successful determination in accomplishing goals in the past, present and future. Hope Pathway relate to a person's perceived capability in being able to overcome goal-related obstacles and produce successful means to accomplish goals. Both of these components are necessary for hopeful thinking as although hope agency

and hope pathway are complementary and positively related they are not synonymous (Synder et al., 1991).

1. ^b	I can think of many ways to get out of a jam.
2. ^a	I energetically pursue my goals.
3.*	I feel tired most of the time.
4. ^b	There are lots of ways around any problem.
5.*	I am easily downed in an argument.
6. ^b	I can think of many ways to get the things in life that are important to me.
7.*	I worry about my health.
8. ^b	Even when others get discouraged, I know I can find a way to solve the problem.
9. ^a	My past experiences have prepared me well for my future.
10. ^a	I've been pretty successful in life.
11.*	I usually find myself worrying about something.
12. ^a	I meet the goals that I set for myself.

^a Hope Agency Items. ^b Hope Pathway Items. * Filler Items

Table 2.5: Trait Hope Scale

This is a 12 item scale that consists of four agency items, four pathway items and four filler items. Each item is scored on a 4 point scale with responses: 'definitely false' (1), 'mostly false' (2), 'mostly true' (3) and 'definitely true' (4). The scores for hope agency and hope pathway are calculated by taking the sum of the 4 scores giving a score range of between 4 and 16. To calculate an overall hope score the sum of hope agency and hope pathway are taken giving a score range of between 8 and 32. The four filler items are included as distracters to break the response sets.

The higher the agency score the more sense of successful determination a person has in relation to the achieving goals generally. The higher the pathway score the more a belief a person has in being able to produce routes in achieving their goals. A high hope score of succeeding in reaching goals cannot be achieved without both a high hope agency score and high hope pathway score.

2.1.2 Data Variables

Each of the variables used within the study was split into groups based on guidelines from a previous study and coded for subsequent analysis. For Gender, Females were coded as 0 and Males coded as 1.

Age was split into two categories, “Mature” coded as 0 and “Under” coded as 1. For undergraduate students, students aged 21 and over on the first day of Session 2009-2010 were placed into the “Mature” category and then those under 21 were “Under”; for postgraduate students those who were aged 25 and over were classed as “Mature” and those under 25 were “Under”.

Domicile was categorized as “Scotland”, “Rest of the UK”, “Rest of Europe” and “Rest of the World”; these were coded as 0, 1, 2 and 3 respectively. Domicile was chosen over nationality since this is what the university uses to see if the student is classed as an overseas student or not and also the student is more likely to have picked up the culture of the domicile status since they have resided there for many years.

Faculty was classified as “Non-Profession” and “Profession”, respectively coded as 0 and 1. Students in Medicine, Veterinary Medicine, Dentistry, Law and Accountancy were deemed “Profession” and the rest of students were deemed “Non-profession”. The reason behind this classification was due to the following differences between the two classes:

- The entry tariffs for “Profession” courses are higher and more competitive than “Non-Profession” students.
- Students in a “Profession” faculty usually have a fixed curriculum leading to cohesive student groups, whereas students in a “Non-profession” faculty have a considerable course choice and may not encounter the same peers in more than one course.

Social Economic Class (SEC) was re-classed into 3 groups “A” coded as 0, “B” coded as 1 and “C” coded as 2. NS-SEC groups 1 and 2 were classed as A, group 3 as B and groups 4, 5, 6, 7, 8 together as C. Groups 0 and 9 were either missing data or not applicable. Since SEC was obtained based on the student’s parents’ occupation (self-reported) it was not recorded for postgraduate students because it is expected that postgraduate students are no longer solely dependent on their parents.

Continuation was split into two categories “Yes” coded as 1 and “No” coded as 0. This is based on whether a student did or did not register at the University of Glasgow the following session (Session 2010-2011), regardless of whether or not the student advanced on at university or repeated the year.

Progression was also split into two categories “Yes” coded as 1 and “No” coded as 0. This is determined by whether a student has progressed to the next year of their original (or cognate) degree programme or not.

2.2 Missing Data

In this study there is a variety of reasons as to how missing data has arisen. Students have occasionally missed out items while other items have been attempted within the same psychometric scale leading to no score being calculated for that psychometric scale. There are also cases where students have missed out whole scale items although they had completed other scales. Possible reasons for why this has happened are that students found the questions too embarrassing or invasive; students may not have understood the question; there could be cross cultural differences for foreign student. The key concepts of missing data and methods by which to deal with missing data will be discussed in this section.

2.2.1 Missing Data Mechanisms

When analysing datasets with missing values it is extremely valuable to establish the nature of the mechanism by which the missing data may have arisen and whether or not the missingness is linked to the underlying values of the variables in the dataset. There are three types of missing data mechanisms: Missing Completely at Random (MCAR), Missing at Random (MAR) and Not Missing at Random (NMAR). It is highly important to establish which missing data mechanism might be at work as the appropriate statistical methods used to analyse the data depend strongly on this. If the manner in which the missing data has arisen is ignored, the results of the statistical methods used may be biased or produce invalid inferences.

The concept of missing data mechanisms was first introduced by Rubin in 1976 where missing data indicators were treated as random variables and a distribution was assigned to

them. This theory is now in common use throughout the modern area of missing data, although the notation and terminology differ slightly from that in the original paper.

Using Little and Rubin's (2002) notation, suppose $Y = (y_{ij})$ is an $n \times k$ rectangular dataset and y_{ij} is the value of the variable Y_j for subject i . Now consider Y has some elements that contain missing values then Y can be written as $Y = (Y_{obs}; Y_{mis})$ where Y_{obs} relates to all the observed entries in Y and Y_{mis} to the missing components. Then let $Y = (Y_{obs}; Y_{mis})$ be the complete data set.

Define the matrix $M = (m_{ij})$ as the missing-data indicator matrix where the number of entries of M matches the number of entries of Y . Let $m_{ij} = 1$ if y_{ij} is missing and $m_{ij} = 0$ if y_{ij} is present.

The missing data mechanism is determined by the conditional distribution of M given $Y = (Y_{obs}; Y_{mis})$, say $f(M | Y, \phi)$, where ϕ denotes the unknown parameters that characterize the relationship between Y and M .

Missing data are termed Missing Completely At Random (MCAR) if the probability of the data being missing is independent of the value of the Y , observed or missing. This is the most restrictive assumption, which can be written as:

$$f(M | Y, \phi) = f(M | \phi) \text{ for all } Y, \phi \quad (2.1)$$

Under the MCAR mechanism the missing data are considered missing completely at random and subjects with missing data can be considered as a random selection of the sample of data. Valid inferences can therefore be made using the non-missing values; the observed sample remains an unbiased representation of the original population (Kenward and Carpenter, 2007).

The second missing data mechanism, Missing at Random (MAR), is less restrictive than the MCAR assumption. In a MAR mechanism, there is a relationship between M and Y_{obs} , but not between M and Y_{mis} . Hence, the missingness depends on the values of the observed data, but not the values of the missing data. This mechanism can be stated as

$$f(M | Y, \phi) = f(M | Y_{obs}, \phi) \text{ for all } Y_{mis}, \phi \quad (2.2)$$

This assumption is key to many analyses with missing data. The MAR assumption does not suggest that the data values are a random sample of all data values (as under MCAR) but requires only that the missing values behave like a random sample of all values within subclasses defined by the observed data (Schafer, 1997).

The last missing data mechanism is called Not Missing at Random (NMAR). In a NMAR mechanism, there is a relationship between M and Y_{mis} , so the missingness depends on the missing values in Y . NMAR is often referred to as non-ignorable missingness since the probability of missing data is related to at least some elements of Y_{mis} and the missing data mechanism cannot be ignored. As a result of this future unobserved responses cannot be predicted. Valid inferences are only possible if the missing data mechanism can be incorporated into the analysis.

2.2.2 Method of Handling Missing Data

There are various methods in which missing data can be dealt with so that eventually standard complete data statistical analysis can be applied.

2.2.2.1 Complete Case Analysis

A common technique to account for missing data is to include only those cases for which all measurements required for a piece of analysis have been observed; this is known as Complete Case Analysis. This is a simple and easy method to employ. However, Diggle et al. (2002) deem it as an “inadequate solution to the problem”.

As all cases with missing values are omitted it can result in a very substantial loss of information, and this gives an impact on reduced statistical precision and power. The conceivable loss of information in removing the incomplete cases from the analysis is a disadvantage of Complete Case analysis. Little and Rubin (2002, p.41) mention the following disadvantages of a Complete Case analysis: the observations with no missing values may not represent the intended study population of interest and there is a loss of precision and an increase in bias when MCAR is not the missing data mechanism. For this reason, this method

is only viable in MCAR settings when the fraction of observations with missing values is small and there is a very large sample size relative to a small portion of missing information.

The advantages of using this method are that it is simple and easy to implement and that standard complete data statistical analysis can be applied without needing any data structure adjustments. Also if the assumption of MCAR holds then it can produce unbiased estimates for the parameters.

This a common technique to deal with the personal attribute scales in section 2.1. Another common technique when dealing with the personal attribute scales is that studies will calculate the mean for each personal attribute scale rather than disregarding all the data from individuals with missing values. Schwarz (2011) states that when “no more than three items on the ten-item” self efficacy scale are missing for a subject then the mean of the non-missing items should be calculated and used. However, this can lead to bias. For this reason, this approach of calculating the mean did not seem a sensible approach to take and was not used in this analysis.

2.2.2.2 Imputation

Another frequently used technique to account for missing data is imputation where the missing values are filled (imputed) in, usually using the observed values that are available. Single imputation is when the missing values are filled in once and multiple imputation is when the missing values are filled in 2 or more times. Imputation procedures produce complete datasets so that analysis conducted on the dataset(s) makes more effective use of all of the observed data.

A simple method of imputation is mean imputation where the missing value of a predictor variable is imputed with the mean of the observed values for that variable. Although this method is simple to perform, the disadvantage is that no additional information is being added as the overall mean will be identical whether the missing values have been imputed or not. Other disadvantages are that the distribution for these variables can be severely distorted, leading to the standard deviations being underestimated. The assumption of MCAR is assumed to be valid for this method.

Regression imputation is also a method of imputation often used. This is where regression is used to predict values for the missing entries of a variable based on other variables that have been measured for the subjects in the study. As other information observed on a subject is taken in account when imputing a value for that subject this makes regression imputation a better choice than mean imputation. To avoid underestimating standard errors, a random variation can be added to each missing case. This allows for fluctuations in the data from the regression line to help solve the problem of underestimated standard errors (Gelman & Hill 2006).

Another well known imputation technique is hot deck imputation. In this method missing values are imputed with values from similar responding units in the sample. This method of imputation is very common in survey settings and can involve complex schemes for selecting subjects that are “similar” for imputation purposes (Little and Rubin, 2002).

The main disadvantages of single imputation is that imputing a single value treats that value as known, and thus, without special adjustments, single imputation can not reflect sampling variability under one model for non-response or uncertainty about the correct model for non-response (Little and Rubin, 2002). Furthermore, inferences about parameters based on filled-in data do not account for imputation uncertainty and will result in underestimated standard errors and confidence intervals that are too narrow.

Multiple imputation has become “an important and influential approach for dealing with the statistical analysis of incomplete data” (Molenberghs & Kenward 2007) since the concept was introduced by Rubin (1976).

Multiple imputation is a technique that involves filling-in missing data repeatedly to create a set of $D \geq 2$ complete datasets. The datasets are subsequently analysed using standard methodology and the results of each set of analyses combined. Multiple imputation assumes that the underlying missing data mechanism at work is MAR. If the MAR assumption does not hold and the missing data mechanism at work is NMAR then this can result in biased estimates.

To create the D completed datasets, a single imputation method such as regression imputation can be used and repeated D number of times in order to create the multiple datasets. After the D completed datasets have been created, the standard statistical analysis is applied to all D datasets to produce D different sets of the complete data estimate and the associated variance for

an estimated parameter \mathcal{G} . The results are then combined using the formulae below (Little and Rubin, 2002):

Let $\hat{\theta}_d$ and W_d denote the parameter and variance estimates of \mathcal{G} , respectively, from the multiply imputed data sets $d = 1, 2, \dots, D$.

The combined estimate from the D multiple datasets is:

$$\bar{\mathcal{G}}_D = \frac{1}{D} \sum_{d=1}^D \hat{\mathcal{G}}_d \quad (2.3)$$

The associated variance estimate of $\bar{\mathcal{G}}_D$ consists of two components known as the average within-imputation variance,

$$\bar{W}_D = \frac{1}{D} \sum_{d=1}^D W_d, \quad (2.4)$$

and the between-imputation variance,

$$B_D = \frac{1}{D-1} \sum_{d=1}^D (\hat{\mathcal{G}}_d - \bar{\mathcal{G}}_D)^2. \quad (2.5)$$

The total variability of $\bar{\mathcal{G}}_D$ is then defined as

$$T_D = \bar{W}_D + \frac{D+1}{D} B_D \quad (2.6)$$

where $(1 + 1/D)$ is an adjustment for a finite number of multiple imputed data sets.

The aim of Multiple Imputation is to replicate the variability that naturally occurs in the data and incorporates uncertainty arising from the imputation process. The variance information provides information about this variability and indicates how the method performs.

2.2.2.3 EM Algorithm

Dempster, Laird and Rubin (1977) introduced the Expectation maximization (EM) algorithm as an iterative algorithm which is used to calculate maximum likelihood estimates in parametric models for incomplete data. As with the Single and Multiple Imputation procedures, the EM Algorithm approach assumes that the missing data are Missing at Random. So, the observed data can be used in some way, or another, to fill in values for the missing data. The EM algorithm follows the process of replacing each missing value by estimated values, then estimating the parameters, then re-estimating the missing values using the new, assumed correct, parameter estimates and then the parameters are re-estimated. This process continues until convergence has been reached.

The EM algorithm is an iterative procedure where each of the iterations consists of 2 steps: the expectation (E) step and the maximization (M) step. The E step involves computing the conditional expectation of the complete data log-likelihood given the observed data and the parameter estimates, $E[l(\theta|Y)|Y_{obs}, \theta^{(t)}]$. The M step is found by maximizing the complete data log-likelihood from the E-step to obtain the parameter estimates. Iteration between the E and M steps occurs until convergence. Convergence is found when the difference between two iterations is arbitrarily small. A disadvantage of the EM algorithm is that when the amount of missing data is large, the rate of convergence can be very slow. However it can be shown that when it does converge it converges reliably, in a manner that it converges to a local maximum or saddle point of the likelihood. It is also conceptually easy to construct and simple to program.

2.3 Methods of Analysis

Different statistical methods used throughout this thesis are described below.

2.3.1 Fisher's Exact Test

To determine if there are associations between two categorical variables Fisher's Exact test will be used. In this thesis it is used to test the association between an indicator variable recording whether or not a personal attribute score could be calculated and each of the demographic variables. It is also used to test the association between an indicator variable recording whether or not a questionnaire was completed and each of the demographic

variables. This was chosen rather than a Chi-squared test since the numbers of missing values are very small for some combinations of the demographic variables, casting doubt on the assumptions that underpin that test.

Using Weisstien's (MathWorld) notation, let there be two categorical variables, X with m categories and Y with n categories. The data can be summarised in an $m \times n$ table Z

$$Z = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

where a_{ij} is the number of observations where $x = i$ and $y = j$. Let R_i be the sum of i th row and C_j be the sum of the j th column and $N = \sum_{i=1}^m R_i = \sum_{j=1}^n C_j$ be the total sum of Z . The conditional probability of getting the actual observed values of table Z given the particular row and column sums is

$$P_{cutoff} = \frac{(R_1!R_2!\cdots R_m!)(C_1!C_2!\cdots C_n!)}{N! \prod_{ij} a_{ij}!} \quad (2.7)$$

This is a multivariate generalization of the hypergeometric probability function. The next step is to calculate P_{cutoff} for all possible tables where R_i and C_j is equal to R_i and C_j for observed table Z . The sum of these probabilities must be 1. The p-value of table Z is calculated by summing the P_{cutoff} for possible tables where P_{cutoff} is less than equal to P_{cutoff} for the observed table Z .

2.3.2 Binary Logistic Regression

Binary logistic regression is a form of generalized linear model that is used for binomial regression. The outcome variable Y is a binary random variable, and depends on one or more explanatory variables $\mathbf{x} = (1, x_1, x_2, \dots, x_p)$, which can be continuous variables or categorical variables. The outcome variable Y_i has two levels with $Y_i = 1$ (response) and $Y_i = 0$ (non-response) with probabilities $P(Y_i = 1 | \mathbf{x}) = \pi(\mathbf{x})$ and $P(Y_i = 0 | \mathbf{x}) = 1 - \pi(\mathbf{x})$. The form of the logistic regression model is:

$$\pi(\mathbf{x}) = \frac{e^{\beta^T \mathbf{x}}}{1 + e^{\beta^T \mathbf{x}}} \quad (2.8)$$

where

$$\beta^T \mathbf{x} = \log\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p \quad (2.9)$$

$\beta^T \mathbf{x}$ is the logit transformation which describes the log odds of $Y_i = 1$ as a linear function of the explanatory variables.

Throughout the thesis the following binary logistic regressions shall be modelled:

- Completion - $Y_i = 1$ if a student completes the questionnaire at baseline and $Y_i = 0$ if a student does not complete the questionnaire at baseline.
- Continuation - $Y_i = 1$ if a 1st year student continues at the University of Glasgow after 1st year and $Y_i = 0$ if a 1st year student does not continue at the University of Glasgow after 1st year.
- Progression - $Y_i = 1$ if a 1st year student progresses their original degree program at the University of Glasgow after 1st year and $Y_i = 0$ if a 1st year student does not progress their original degree program at the University of Glasgow after 1st year.

2.3.3 Model Building

When modelling the outcome variable Y , there are a number of explanatory variables that could potentially significantly contribute to the outcome variable. To determine the model that best describes the relationship between the outcome variable and the explanatory variables the process of model building shall be used where explanatory variables are added and dropped from the model. To compare the models the same data set must be used for every model, therefore any observations with missing explanatory variables must be removed from the dataset. The methods used to determine which model is best to describe the outcome variable are the Deviance using Generalized Likelihood Ratio Test, AIC and BIC.

The Generalized likelihood ratio test (GLRT) compares two model deviances, denoted D_i . Models are compared in a hierarchical method of selecting or eliminating particular explanatory variables from the model. The GLRT statistic G (Hosmer and Lemeshow 1989) is the difference in deviance between model 2 (model missing additional variable β_t) and model 1 (model with additional variable β_t)

$$G = D_2 - D_1$$

and is compared to a chi-squared distribution

$$\chi^2(p_1 - p_2)$$

where p_1 is the number of parameters in model 1 and p_2 is the number of parameters in model 2. The null hypothesis is the slope coefficients of the additional variable $\beta_t = 0$.

The null hypothesis is rejected if:

$$D_2 - D_1 > \chi^2(p_1 - p_2; 1 - \alpha) \quad (2.10)$$

for a test of size approximately α .

AIC (Akaike's Information Criterion) is computed for every possible type of model and calculated as:

$$AIC = Deviance + 2p \quad (2.11)$$

where p is the number of parameters in the model. The AIC determines if extra parameters in the model are justified by penalising the deviance of the model. The best model is determined as the model with the smallest AIC value (Akaike, 1978).

Throughout this thesis the smallest AIC will be used to determine the best model. This method may not always be used, instead further restrictions are applied to AIC (Burnham and Anderson, 2004). However, it is widely believed that having penalised the likelihood already, in order to obtain values of AIC, it is not appropriate to apply further rules to restrict the choice of best model.

BIC (Bayesian Information Criterion) is very similar to the AIC except that it also takes sample size, denoted n , into account. It is calculated as:

$$BIC = Deviance + p \log(n) \quad (2.12)$$

Again, the best model is determined as the model with the smallest BIC value.

2.3.4 Hosmer-Lemeshow Test

The Hosmer-Lemeshow test is a test that assesses the goodness of fit of a logistic regression model. The null hypothesis is that the model is an adequate fit to the data while the alternative hypothesis is that the model is not an adequate fit to the data. Hosmer and Lemeshow (1989) proposed a test statistic that they show, through simulation, is distributed approximately as chi-square under the null hypothesis, when there is no replication in any of the subgroups defined by combinations of the explanatory variables. (For example, if the logistic regression model had two binary explanatory variables, x_1 and x_2 , then there would be 4 subgroups of cases defined by the four possible combinations of the levels of x_1 and x_2 .)

To begin with the observations are sorted in increasing order of their estimated event probability then the observations are partitioned into G equal sized groups (where G is usually about 10). The Hosmer-Lemeshow goodness-of-fit statistic is obtained by calculating the Pearson chi-square statistic from the $2 \times G$ table of observed and expected frequencies, where G is the number of groups. The statistic is written as

$$H_L = \sum_{g=1}^G \frac{(O_g - N_g \bar{\pi}_g)^2}{N_g \bar{\pi}_g (1 - \bar{\pi}_g)} \sim \chi^2 (G - 2) \quad (2.13)$$

where N_g is the total frequency of subjects in the g -th group, O_g is the total frequency of event outcomes in the g -th group, $\bar{\pi}_g$ is the average estimated probability of an event outcome for the g -th group. The test statistic asymptotically follows a χ^2 distribution with $G - 2$ degrees of freedom.

It must be possible to define at least three different groups in order for the Hosmer-Lemeshow statistic to be computed. Therefore the Hosmer-Lemeshow test is not appropriate for models containing only 1 binary explanatory variable.

2.4 Statistical Programs

The analysis for this thesis will use **R**. **R** (R Development Core Team, 2011) is a free and widely used statistical language for statistical computing. The advantages of **R** were that it was free and could be downloaded on to any computer and that the **mi** package was available and fairly flexible to impute the missing data.

To impute data in this thesis function from the **mi package** in **R** have been used (Su et al 2009). The **mi** package uses Iterative EM-based multiple Bayesian regression imputation of missing values. The **mi.info** function is used to produce a matrix of imputation information needed by the **mi** function to impute the missing data. The **mi.info** function extracts information from the dataset and creates default model specifications which can then be updated by the user. This information matrix includes information on such things as the names of the variables, the number of data points missing in each variable, the variable type, whether a variable is to be included in the imputation model or not and the imputation formulas used in the imputation models. The **mi** function is then used to impute the missing data, where the original data frame is stated, as well as the information matrix, the number of imputations, the maximum number of imputation iterations, and whether to check convergence of the coefficients of the imputation models. After this the **write.mi** function is used to write the imputed datasets to a file in csv format.

Chapter 3

Data Description

3.1 Exclusion Criteria

The main interest for this study is the proportion of 1st year students who withdraw from the University of Glasgow during their first year and how their personal attributes and change in personal attributes are linked to early withdrawal. Therefore it was decided to implement exclusion criteria to decide exactly whose responses would be included in the analysis, with the intention of obtaining a true representation of the population of 1st year students.

Before being given the data, the sample size had been reduced from 1545 to 1504 by the research team. The bases for these exclusions were: no consent given; respondent not identifiable from registration number and name supplied; duplicate responses. The respondents not identifiable from the registration number and name supplied were excluded for several reasons. To begin with, if they could not be identified then there would be no way of being able to establish if they withdrew from the university. Another reason is that the information they provided appeared to be incorrect or fake: if they could not fill out the information correctly it would be more than likely that they would not be able to fill out the questionnaire correctly.

Once given the data it was decided that more responses would need to be excluded to obtain a true representation of the population of 1st year students. Students who never fully registered were excluded since they were never formally students of the university. This led to students who had deferred entry also being excluded as they had not officially started university and instead belonged to the following year's cohort.

Visiting students were excluded because they had already had experience of higher education elsewhere and would not be graduating from the University of Glasgow but leaving once their visit was over. It was then decided also to exclude incoming exchange students for the same reason. This then led to students who were abroad for languages to be excluded.

Following these exclusions it was decided to exclude students who were not in 1st year given that our main interest is how the personal attributes of 1st year students are linked to early withdrawal. These are students who may have gone straight into 2nd, 3rd or 4th year instead of starting at 1st year.

In view of this it was also decided to exclude undergraduate students who had already obtained a degree before entering another course of study. This is because it would not be their first experience of being an undergraduate so had already been subjected to how university life is.

Although it would have been preferable, unfortunately it was not possible to exclude all undergraduate students who may have previously started a degree but whose credit did not count towards entry into their current degree at the University of Glasgow. This was decided because, although we know how many students decided to restart at Glasgow from their matriculation number (7 students), there is no way of knowing how many students may have attended another university then entered a new degree programme at Glasgow with no credit from their prior higher education study.

Part time students and distance learning students were not excluded since we are looking at how the personal attributes change as a result of exposure to university, not so much the setting.

After applying the above exclusion criteria we obtain a sample size of 1373: 969 undergraduate students and 404 postgraduate students.

The number of responses excluded at each stage of the exclusion criteria can be seen in Table 3.1.

Number of Students Excluded				Sample Size after Exclusion		
UG	PG	Total	Justification	UG	PG	Total
						1545
n/a	n/a	7	No Consent & No Response	n/a	n/a	1538
n/a	n/a	20	Respondent Not Identifiable from Registration No. and Name Supplied	n/a	n/a	1518
n/a	n/a	14	Duplicate Response	1096	408	1504
2	0	2	Never Fully Registered	1094	408	1502
2	0	2	Deferred Entry	1092	408	1500
50	3	53	Visiting Students	1042	405	1447
7	0	7	Incoming Exchange	1035	405	1440
1	0	1	Language Abroad	1034	405	1439
31	1	32	Students Not in 1 st Year	1003	404	1407
34	0	34	Undergraduate Students With A Degree	969	404	1373
127	4	172	Total	969	404	1373

Table 3.1 Table of Excluded Data

3.2 Descriptive Statistics

Table 3.2 and Table 3.4 show the number of students in each of the demographic groups at Baseline, Semester 1 and Semester 2 for Undergraduates and Postgraduates that were obtained after the exclusion criteria had been applied. It is not possible to establish if the sample demographic groups are representative of the whole university cohort as this information was not available. Table 3.3 and Table 3.5 show the median, lower quartile and upper quartile of each of the personal attribute scores that could be calculated at Baseline, Semester 1 and Semester 2 for Undergraduates and Postgraduates.

Demographic Variables		Baseline (N=969)		Semester 1 (N=193)		Semester 2 (N=152)	
Sex	Female	604	(62.33%)	133	(68.91%)	101	(66.45%)
	Male	365	(37.67%)	60	(31.09%)	51	(33.55%)
Age	Mature	134	(13.83%)	34	(17.62%)	32	(21.05%)
	Under 21	835	(86.17%)	159	(82.38%)	120	(78.95%)
Faculty	Non Profession	794	(81.94%)	153	(79.27%)	124	(81.58%)
	Profession	175	(18.06%)	40	(20.73%)	28	(18.42%)
Domicile	Scotland	675	(69.66%)	129	(66.84%)	108	(71.06%)
	Rest of the UK	142	(14.65%)	34	(17.62%)	19	(12.50%)
	Rest of Europe	113	(11.66%)	25	(12.95%)	17	(11.18%)
	Rest of the World	36	(3.72%)	5	(2.59%)	8	(5.26%)
	Unknown	3	(0.31%)	0	(0.00%)	0	(0.00%)
SEC	A	409	(42.21%)	81	(41.97%)	59	(38.82%)
	B	93	(9.60%)	20	(10.36%)	13	(8.55%)
	C	129	(13.31%)	24	(12.44%)	20	(13.16%)
	Unknown	338	(34.88%)	68	(35.23%)	60	(39.47%)

Table 3.2: Demographic Variables by Time Point for Undergraduates

Personal Attribute (Scale Range)	Baseline			Semester 1			Semester 2		
	Median	Q1	Q3	Median	Q1	Q3	Median	Q1	Q3
Mindset (1 to 6)	3.50	3.00	4.25	3.25	2.50	4.00	3.50	2.75	4.00
Self Efficacy (10 to 40)	31.00	29.00	34.00	30.00	29.00	33.00	31.00	29.00	33.00
Self Esteem (10 to 40)	31.00	28.00	34.00	30.00	27.00	33.00	30.00	27.00	33.00
Resilience (1 to 4)	3.00	2.80	3.20	2.90	2.70	3.10	2.90	2.60	3.10
Hope Total (8 to 32)	25.00	24.00	27.00	24.00	22.75	26.00	25.00	23.00	26.00
Hope Agency (4 to 16)	13.00	12.00	14.00	12.00	11.00	13.00	12.00	12.00	13.00
Hope Pathway (4 to 16)	12.00	12.00	13.00	12.00	11.00	13.00	12.00	11.00	13.00

Table 3.3: Personal Attributes by Time Point for Undergraduates

Demographic Variables		Baseline (N=404)		Semester 1 (N=93)		Semester 2 (N=78)	
Sex	Female	248	(61.39%)	58	(62.37%)	52	(66.67%)
	Male	156	(38.61%)	35	(37.63%)	26	(33.33%)
Age	Mature	235	(58.17%)	53	(56.99%)	45	(57.69%)
	Under 25	169	(41.83%)	40	(43.01%)	33	(42.31%)
Faculty	Non Profession	244	(60.40%)	53	(56.99%)	46	(58.97%)
	Profession	160	(39.60%)	40	(43.01%)	32	(41.03%)
Domicile	Scotland	191	(47.28%)	40	(43.01%)	44	(56.41%)
	Rest of the UK	25	(6.19%)	7	(7.53%)	4	(5.13%)
	Rest of Europe	59	(14.60%)	10	(10.75%)	10	(12.82%)
	Rest of the World	121	(29.95%)	32	(34.41%)	20	(25.64%)
	Unknown	8	(1.98%)	4	(4.30%)	0	(0.00%)

Table 3.4: Demographic Variables by Time Point for Postgraduates

Personal Attribute (Scale Range)	Baseline			Semester 1			Semester 2		
	Median	Q1	Q3	Median	Q1	Q3	Median	Q1	Q3
Mindset (1 to 6)	4.00	3.25	4.75	3.50	2.75	4.00	3.75	3.00	4.00
Self Efficacy (10 to 40)	32.00	29.00	35.00	32.00	30.00	35.00	31.00	28.00	35.00
Self Esteem (10 to 40)	31.00	28.00	34.00	31.00	28.00	35.00	30.00	27.00	33.00
Resilience (1 to 4)	2.90	2.80	3.10	3.00	2.80	3.20	2.90	2.60	3.10
Hope Total (8 to 32)	25.00	24.00	27.00	25.00	24.00	27.00	25.00	22.50	26.00
Hope Agency (4 to 16)	13.00	12.00	14.00	13.00	12.00	14.00	12.00	11.25	13.00
Hope Pathway (4 to 16)	12.00	12.00	13.00	12.00	12.00	13.00	12.00	11.00	13.00

Table 3.5: Demographic Variables by Time Point for Postgraduates

3.3 Non Response

Wave Non-response, for this study, is when a student does not attempt a follow up questionnaire in Semester 1 or Semester 2. Possible reasons for this could be that the student found the Baseline questionnaire personal and invasive or that the timing wasn't convenient when asked to fill out the follow up questionnaire.

	Wave Non-Response	
Semester 1		
UG	282/475	(59.37%)
PG	113/206	(54.85%)
Total	395/680	(58.09%)
Semester 2		
UG	342/494	(69.23%)
PG	120/198	(60.61%)
Total	462/695	(66.47%)
Overall Follow Up		
UG	624/969	(64.40%)
PG	233/404	(57.67%)
Total	857/1373	(62.42%)

Table 3.6: Rate of Wave Non-Response

Table 3.6 provides the percentage of wave non-response for each semester individually and the percentage of wave non-response overall for Undergraduates and Postgraduates.

It can be seen from Table 3.6 that the over half of the Undergraduates and Postgraduates did not attempt a follow up questionnaire in Semester 1 and Semester 2. The percentage of non response for Semester 2 is higher than in Semester 1 for Undergraduates and Postgraduates. Overall 64.40% of the Undergraduates and 57.67% of Postgraduates did not attempt a follow up questionnaire.

In the context of this study, Item Non-response is when a particular question of the questionnaire has been missed out or purposely not been answered by a student. Possible reasons for this could be that the student found the meaning of the question confusing or that the student felt it was too personal and invasive. As explained earlier, Item Non-response leads to a Personal Attribute Scale non-response.

It will be of interest to compare the missingness of each question and personal attribute scale within the questionnaire and also how the missingness of each question and personal attribute scale changes at each time point of the study.

Table 3.7 provides the percentage of non-response for each question and attribute scale for Undergraduates and Postgraduates at Baseline, Semester 1 and Semester 2.

Looking at Table 3.7, it can be seen that the percentage of non-response for each question at Baseline ranges between 0.1% to 1.1% for Undergraduates and between 0.5% and 3.2% for Postgraduates. This percentage increases at Semester 1 and Semester 2. The percentage of non-response for each question within a scale appears to be fairly evenly spread across the questions. This is consistent for Undergraduates and Postgraduates at each time point. This gives the impression that there is no pattern of missingness for item non-response.

The percentage of non-response for each individual personal attribute scale at Baseline ranges from 0.9% to 4.8% for Undergraduates and between 2.7% and 8.7% for Postgraduates. Again this percentage increases at Semester 1 and Semester 2. Self Efficacy and Resilience appear to have higher percentage of non-response consistently for Undergraduates and Postgraduates across each time point.

	Baseline						Semester 1						Semester 2					
Attribute	UG		PG		Total		UG		PG		Total		UG		PG		Total	
Mindset Q1	1	0.1%	3	0.7%	4	0.3%	1	0.5%	1	1.1%	2	0.7%	1	0.7%	0	0.0%	1	0.4%
Mindset Q2	2	0.2%	4	1.0%	6	0.4%	1	0.5%	1	1.1%	2	0.7%	1	0.7%	0	0.0%	1	0.4%
Mindset Q3	2	0.2%	4	1.0%	6	0.4%	2	1.0%	2	2.2%	4	1.4%	2	1.3%	1	1.3%	3	1.3%
Mindset Q4	4	0.4%	4	1.0%	8	0.6%	1	0.5%	1	1.1%	2	0.7%	1	0.7%	0	0.0%	1	0.4%
Mindset	9	0.9%	11	2.7%	20	1.5%	2	1.0%	2	2.2%	4	1.4%	2	1.3%	1	1.3%	3	1.3%
Self Efficacy Q1	1	0.1%	2	0.5%	3	0.2%	1	0.5%	1	1.1%	2	0.7%	3	2.0%	0	0.0%	3	1.3%
Self Efficacy Q2	5	0.5%	3	0.7%	8	0.6%	2	1.0%	0	0.0%	2	0.7%	3	2.0%	1	1.3%	4	1.7%
Self Efficacy Q3	5	0.5%	6	1.5%	11	0.8%	2	1.0%	0	0.0%	2	0.7%	1	0.7%	1	1.3%	2	0.9%
Self Efficacy Q4	8	0.8%	9	2.2%	17	1.2%	3	1.6%	1	1.1%	4	1.4%	2	1.3%	1	1.3%	3	1.3%
Self Efficacy Q5	4	0.4%	6	1.5%	10	0.7%	3	1.6%	0	0.0%	3	1.0%	2	1.3%	1	1.3%	3	1.3%
Self Efficacy Q6	1	0.1%	9	2.2%	10	0.7%	2	1.0%	0	0.0%	2	0.7%	3	2.0%	0	0.0%	3	1.3%
Self Efficacy Q7	11	1.1%	8	2.0%	19	1.4%	2	1.0%	1	1.1%	3	1.0%	1	0.7%	0	0.0%	1	0.4%
Self Efficacy Q8	10	1.0%	5	1.2%	15	1.1%	1	0.5%	0	0.0%	1	0.3%	1	0.7%	0	0.0%	1	0.4%
Self Efficacy Q9	3	0.3%	9	2.2%	12	0.9%	2	1.0%	3	3.2%	5	1.7%	1	0.7%	0	0.0%	1	0.4%
Self Efficacy Q10	2	0.2%	5	1.2%	7	0.5%	2	1.0%	0	0.0%	2	0.7%	1	0.7%	0	0.0%	1	0.4%
Self Efficacy	41	4.2%	29	7.2%	70	5.1%	9	4.7%	4	4.3%	13	4.5%	8	5.3%	4	5.1%	12	5.2%
Self Esteem Q1	1	0.1%	6	1.5%	7	0.5%	1	0.5%	0	0.0%	1	0.3%	1	0.7%	1	1.3%	2	0.9%
Self Esteem Q2	1	0.1%	10	2.5%	11	0.8%	2	1.0%	0	0.0%	2	0.7%	1	0.7%	1	1.3%	2	0.9%
Self Esteem Q3	2	0.2%	7	1.7%	9	0.7%	2	1.0%	1	1.1%	3	1.0%	3	2.0%	2	2.6%	5	2.2%
Self Esteem Q4	3	0.3%	6	1.5%	9	0.7%	2	1.0%	0	0.0%	2	0.7%	1	0.7%	1	1.3%	2	0.9%
Self Esteem Q5	3	0.3%	9	2.2%	12	0.9%	1	0.5%	0	0.0%	1	0.3%	1	0.7%	1	1.3%	2	0.9%
Self Esteem Q6	3	0.3%	11	2.7%	14	1.0%	1	0.5%	1	1.1%	2	0.7%	1	0.7%	1	1.3%	2	0.9%
Self Esteem Q7	4	0.4%	13	3.2%	17	1.2%	2	1.0%	1	1.1%	3	1.0%	1	0.7%	1	1.3%	2	0.9%
Self Esteem Q8	6	0.6%	8	2.0%	14	1.0%	1	0.5%	0	0.0%	1	0.3%	2	1.3%	2	2.6%	4	1.7%
Self Esteem Q9	5	0.5%	8	2.0%	13	0.9%	1	0.5%	0	0.0%	1	0.3%	3	2.0%	1	1.3%	4	1.7%
Self Esteem Q10	4	0.4%	7	1.7%	11	0.8%	1	0.5%	0	0.0%	1	0.3%	1	0.7%	1	1.3%	2	0.9%
Self Esteem	25	2.6%	27	6.7%	52	3.8%	5	2.6%	3	3.2%	8	2.8%	6	3.9%	3	3.8%	9	3.9%
Resilience Q1	1	0.1%	8	2.0%	9	0.7%	1	0.5%	0	0.0%	1	0.3%	1	0.7%	0	0.0%	1	0.4%
Resilience Q2	4	0.4%	9	2.2%	13	0.9%	1	0.5%	1	1.1%	2	0.7%	1	0.7%	1	1.3%	2	0.9%
Resilience Q3	3	0.3%	8	2.0%	11	0.8%	1	0.5%	2	2.2%	3	1.0%	1	0.7%	1	1.3%	2	0.9%
Resilience Q4	7	0.7%	9	2.2%	16	1.2%	2	1.0%	2	2.2%	4	1.4%	1	0.7%	0	0.0%	1	0.4%
Resilience Q5	1	0.1%	8	2.0%	9	0.7%	2	1.0%	0	0.0%	2	0.7%	1	0.7%	0	0.0%	1	0.4%
Resilience Q6	5	0.5%	8	2.0%	13	0.9%	2	1.0%	0	0.0%	2	0.7%	2	1.3%	0	0.0%	2	0.9%
Resilience Q7	3	0.3%	10	2.5%	13	0.9%	1	0.5%	0	0.0%	1	0.3%	1	0.7%	0	0.0%	1	0.4%
Resilience Q8	6	0.6%	9	2.2%	15	1.1%	1	0.5%	0	0.0%	1	0.3%	3	2.0%	0	0.0%	3	1.3%
Resilience Q9	8	0.8%	11	2.7%	19	1.4%	2	1.0%	1	1.1%	3	1.0%	2	1.3%	0	0.0%	2	0.9%
Resilience Q10	6	0.6%	11	2.7%	17	1.2%	1	0.5%	0	0.0%	1	0.3%	3	2.0%	0	0.0%	3	1.3%
Resilience Q11	7	0.7%	13	3.2%	20	1.5%	3	1.6%	0	0.0%	3	1.0%	2	1.3%	1	1.3%	3	1.3%
Resilience Q12	8	0.8%	7	1.7%	15	1.1%	3	1.6%	0	0.0%	3	1.0%	1	0.7%	0	0.0%	1	0.4%
Resilience Q13	8	0.8%	10	2.5%	18	1.3%	1	0.5%	3	3.2%	4	1.4%	2	1.3%	0	0.0%	2	0.9%
Resilience Q14	0	0.0%	7	1.7%	7	0.5%	1	0.5%	0	0.0%	1	0.3%	1	0.7%	0	0.0%	1	0.4%
Resilience	47	4.8%	35	8.7%	82	6.0%	8	4.1%	8	8.6%	16	5.6%	6	3.9%	3	3.8%	9	3.9%
Hope Q1	1	0.1%	5	1.2%	6	0.4%	1	0.5%	0	0.0%	1	0.3%	1	0.7%	0	0.0%	1	0.4%
Hope Q2	3	0.3%	6	1.5%	9	0.7%	2	1.0%	0	0.0%	2	0.7%	1	0.7%	0	0.0%	1	0.4%
Hope Q3*	5	0.5%	7	1.7%	12	0.9%	1	0.5%	0	0.0%	1	0.3%	2	1.3%	0	0.0%	2	0.9%
Hope Q4	7	0.7%	5	1.2%	12	0.9%	2	1.0%	0	0.0%	2	0.7%	2	1.3%	1	1.3%	3	1.3%
Hope Q5*	4	0.4%	5	1.2%	9	0.7%	2	1.0%	0	0.0%	2	0.7%	2	1.3%	0	0.0%	2	0.9%
Hope Q6	6	0.6%	5	1.2%	11	0.8%	2	1.0%	0	0.0%	2	0.7%	1	0.7%	1	1.3%	2	0.9%
Hope Q7*	5	0.5%	4	1.0%	9	0.7%	1	0.5%	0	0.0%	1	0.3%	2	1.3%	0	0.0%	2	0.9%
Hope Q8	5	0.5%	5	1.2%	10	0.7%	2	1.0%	0	0.0%	2	0.7%	2	1.3%	1	1.3%	3	1.3%
Hope Q9	6	0.6%	5	1.2%	11	0.8%	1	0.5%	0	0.0%	1	0.3%	4	2.6%	0	0.0%	4	1.7%
Hope Q10	9	0.9%	5	1.2%	14	1.0%	1	0.5%	1	1.1%	2	0.7%	1	0.7%	0	0.0%	1	0.4%
Hope Q11*	4	0.4%	7	1.7%	11	0.8%	2	1.0%	0	0.0%	2	0.7%	2	1.3%	0	0.0%	2	0.9%
Hope Q12	3	0.3%	5	1.2%	8	0.6%	1	0.5%	0	0.0%	1	0.3%	2	1.3%	0	0.0%	2	0.9%
Hope Total	31	3.2%	12	3.0%	43	3.1%	5	2.6%	1	1.1%	6	2.1%	6	3.9%	3	3.8%	9	3.9%
Hope Agency	18	1.8%	9	2.2%	27	2.0%	2	1.0%	1	1.1%	3	1.0%	5	3.3%	0	0.0%	5	2.2%
Hope Pathway	15	1.5%	7	1.7%	22	1.6%	4	2.1%	0	0.0%	4	1.4%	3	2.0%	3	3.8%	6	2.6%

* The items not used to calculate Hope

Table 3.7: Rate of Item Non-Response & Personal Attribute Scale Non-Response

As the non-response for each question within a scale appears to be fairly evenly spread across the questions, the proportion of scores that could not be calculated for each individual personal attribute scale shall now be examined. Fisher's Exact Test has been applied at a 5% significance level to examine the significance of the association between the completion or non-completion of a personal attribute scale and sex. As a number of tests are being carried out at 5% significance level there is potential false positive results occurring.

Null Hypotheses: The population proportions of personal attribute scores that could not be calculated for Females and Males are equal.

Alternative Hypotheses: The population proportions of personal attribute scores that could not be calculated for Females and Males are not equal.

Personal Attribute	No		Yes		Odds Ratio (95% CI)	Fisher's Exact Test p-value
Mindset						
Female	6/604	(0.99%)	598/604	(99.01%)		
Male	3/365	(0.82%)	362/365	(99.18%)	1.21 (0.26, 7.53)	≈ 1
Self Efficacy						
Female	26/604	(4.30%)	598/604	(95.70%)		
Male	15/365	(4.11%)	362/365	(95.89%)	1.05 (0.53, 2.17)	≈ 1
Self Esteem						
Female	14/604	(2.32%)	598/604	(97.68%)		
Male	11/365	(3.01%)	362/365	(96.88%)	0.76 (0.32, 1.88)	0.534
Resilience						
Female	22/604	(3.64%)	598/604	(96.36%)		
Male	25/365	(6.85%)	362/365	(93.15%)	0.51 (0.27, 0.97)	0.030
Hope Total						
Female	21/604	(3.48%)	598/604	(96.52%)		
Male	10/365	(2.74%)	362/365	(97.26%)	1.28 (0.57, 3.08)	0.577
Hope Agency						
Female	11/604	(1.82%)	598/604	(98.18%)		
Male	7/365	(1.92%)	362/365	(98.08%)	0.95 (0.33, 2.91)	≈ 1
Hope Pathway						
Female	12/604	(1.99%)	598/604	(98.01%)		
Male	3/365	(0.82%)	362/365	(99.18%)	2.44 (0.65, 13.59)	0.187

Table 3.8: Fisher's Exact Test of Completed Personal Attribute Scales by Sex at Baseline for Undergraduates

Personal Attribute	No		Yes		Odds Ratio (95% CI)	Fisher's Exact Test p-value
Mindset						
Female	8/248	(3.23%)	240/248	(96.77%)		
Male	3/156	(1.92%)	153/156	(98.08%)	1.70 (0.40, 10.09)	0.541
Self Efficacy						
Female	17/248	(6.85%)	231/248	(93.15%)		
Male	12/156	(7.69%)	144/156	(92.31%)	0.88 (0.38, 2.09)	0.843
Self Esteem						
Female	19/248	(7.66%)	229/248	(92.34%)		
Male	8/156	(5.13%)	148/156	(94.87%)	1.53 (0.62, 4.16)	0.414
Resilience						
Female	22/248	(8.87%)	226/248	(91.13%)		
Male	13/156	(8.33%)	143/156	(91.67%)	1.07 (0.50, 2.39)	≈ 1
Hope Total						
Female	8/248	(3.23%)	240/248	(96.77%)		
Male	4/156	(2.56%)	152/156	(97.44%)	1.27 (0.33, 5.84)	0.773
Hope Agency						
Female	6/248	(2.42%)	242/248	(97.58%)		
Male	3/156	(1.92%)	153/156	(98.08%)	1.26 (0.27, 7.92)	≈ 1
Hope Pathway						
Female	5/248	(2.02%)	243/248	(97.98%)		
Male	2/156	(1.28%)	154/156	(98.72%)	1.58 (0.26, 16.82)	0.771

Table 3.9: Fisher's Exact Test of Completed Personal Attribute Scales by Sex at Baseline for Postgraduates

Table 3.8 and Table 3.9 display the sample population percentages and counts of the number of people whose attribute score could not be calculated for each personal attribute by sex at Baseline for Undergraduates. It can be seen that for every attribute scale, except resilience for undergraduates, the p-value is greater than our significance level of 0.05, therefore we cannot reject our null hypothesis. Hence there is no statistically significant difference in the proportion of personal attribute scales that could not be calculated between the population of males and females.

For resilience we have a p-value of 0.03 for the undergraduate students, which is less than our significance level of 0.05. Therefore we can reject our null hypotheses and state that there is a statistically significant difference between the population proportion of resilience scales that could not be calculated for males and females. A higher proportion of male than female undergraduates failed to complete the Resilience scale.

For Undergraduates and Postgraduates at Semester 1 and 2, the Fisher's Exact Test showed that there was no statistically significant difference in the proportion of attribute scales that could not be calculated between the population of males and females.

Fisher's Exact Test was again used to examine the significance of the association between the ability to calculate the various personal attribute scores which could not be calculated and Age, Faculty and Domicile individually for undergraduate and postgraduate students. For each of these demographic variables the p-value is greater than our significance level of 0.05, therefore we cannot reject our null hypothesis. Hence there is no statistically significant difference in the proportion of personal attribute scales that could not be calculated between the sub-populations defined by each demographic variable individually.

As there appears to be no pattern to the proportion of individual attribute scale scores that could not be calculated, the proportion of questionnaires that were not completed shall be investigated. A binary variable was created with value 1 if a subject answered every question on the questionnaire and a value 0 otherwise. This greatly reduces the number of tests carried out and greatly reduces the chances of false positive results.

Fisher's Exact Test has been applied to examine the significance of the association between a questionnaire not being completed and the demographic variables (Sex, Age, Domicile and SEC) of respondents.

Demographic Variable	No	Yes	Odds Ratio 95% CI	Fisher's Exact Test p-value
Sex				
Female	67/604 (11.09%)	537/604 (88.91%)		
Male	53/365 (14.52%)	312/365 (85.48%)	0.73 (0.49, 1.10)	0.131
Age				
Mature	24/134 (17.91%)	110/134 (82.09%)		
Under 21	96/835 (11.50%)	739/835 (88.50%)	1.67 (0.98, 2.79)	0.047
Faculty				
Non Profession	109/794 (13.73%)	685/794 (86.27%)		
Profession	11/175 (6.29%)	164/175 (93.71%)	2.37 (1.24, 5.00)	0.005
Domicile				
Scotland	86/675 (12.74%)	589/675 (87.26%)		
Rest of the UK	14/142 (9.86%)	128/142 (90.14%)	1.33 (0.74, 2.42)	
Rest of Europe	15/113 (13.27%)	98/113 (86.73%)	0.95 (0.53, 1.72)	
Rest of the World	5/36 (13.89%)	31/36 (86.11%)	0.91 (0.34, 2.39)	0.767
SEC				
A	43/409 (10.51%)	366/409 (89.49%)		
B	13/93 (13.98%)	80/93 (86.02%)	0.72 (0.37, 1.41)	
C	13/129 (10.00%)	116/129 (90.00%)	1.05 (0.54, 2.02)	0.608

Table 3.10: Fisher's Exact Test of Completed Questionnaires by Demographic variables at Baseline for Undergraduates

Demographic Variable	No		Yes		Odds Ratio 95% CI	Fisher's Exact Test p-value
Sex						
Female	49/248	(19.76%)	199/248	(80.24%)		
Male	28/156	(17.95%)	128/156	(82.05%)	1.13 (0.65, 1.96)	0.698
Age						
Mature	50/235	(21.28%)	185/235	(78.72%)		
Under 25	27/169	(15.98%)	142/169	(84.02%)	1.42 (0.83, 2.48)	0.200
Faculty						
Non Profession	50/244	(20.49%)	194/244	(79.51%)		
Profession	27/160	(16.88%)	133/160	(83.12%)	1.27 (0.44, 2.22)	0.437
Domicile						
Scotland	37/59	(19.37%)	154/59	(80.63%)		
Rest of the UK	4/25	(16.00%)	21/25	(84.00%)	1.26 (0.41, 3.90)	
Rest of Europe	11/121	(18.64%)	48/121	(81.36%)	1.05 (0.50, 2.21)	
Rest of the World	23/191	(19.01%)	98/191	(80.99%)	1.02 (0.57, 1.83)	0.996

Table 3.11: Fisher's Exact Test of Completed Questionnaires by Demographic variables at Baseline for Postgraduates

Table 3.10 and Table 3.11 show that every demographic variable, except age and faculty at baseline for undergraduates, have p-values that are greater than our significance level of 0.05, therefore we cannot reject our null hypothesis that the proportions of questionnaires that were not completed are equal within the sub-populations defined by the demographic variables separately.

For age and faculty, respectively, we have a p-value of 0.047 and 0.005 for the undergraduate students, which is less than our significance level of 0.05. Therefore we can reject our null hypothesis and state that there is a statistically significant difference between the population proportions of questionnaires that were not completed for students under 21 and students over 21 and that there is a statistically significant difference between the population proportions of questionnaires that were not completed for students in a non profession faculty and students in a profession faculty. Older undergraduates and those in professional courses are more likely to complete the entire questionnaire.

For Undergraduates and Postgraduates at Semester 1 and 2, the Fisher's Exact Test showed that there was no statistically significant difference in the proportion of questionnaires that were completed and not completed between the sub-populations defined by the demographic variables separately. It was decided not to investigate in further detail the missingness in Semester 1 and 2 as there is little power in the tests because of the small number of responses and the even smaller number of missing responses for Undergraduates and Postgraduates at these time points. Also the Fisher's Exact Tests were not statistically significant for proportion of completed questionnaires and for the proportion of completed personal attribute scale scores.

To further investigate any differences between the sub-population defined by the demographic variables and the completion of questionnaires binary logistic regression models were fitted.

Univariate Logistic Regression was used to model the log odds of questionnaire completion by explanatory variables Age, Sex, Domicile, Faculty and SEC separately. Then all explanatory variables were included in the logistic regression analysis to determine if any are significantly related to whether a questionnaire is fully completed when other explanatory variables are also included in the model. This was done for undergraduates and postgraduates separately.

Demographic Variable	Coef	Std Error	P-value	df	Residual Deviance
Sex					
Intercept	2.081	0.130	<0.001		
Sex (Males)	-0.309	0.197	0.117	967	723.37
Age					
Intercept	1.522	0.225	<0.001		
Age (Under 21)	0.519	0.250	0.038	967	721.79
Faculty					
Intercept	1.838	0.103	<0.001		
Faculty (Profession)	0.864	0.328	0.008	967	721.36
Domicile					
Intercept	1.924	0.115	<0.001		
Domicile (Rest of the UK)	0.289	0.304	0.342		
Domicile (Rest of Europe)	-0.047	0.300	0.875		
Domicile (Rest of the World)	-0.099	0.496	0.841	962	723.88
SEC					
Intercept	2.141	0.161	<0.001		
SEC (B)	-0.324	0.340	0.340		
SEC (C)	0.472	0.334	0.888	628	434.59

Table 3.12: Univariate Logistic Regression of Completion by Demographic Variables at Baseline for Undergraduates

Table 3.12 shows that Age and Faculty separately both have a p-value less than our significance level of 0.05, therefore Age and Faculty separately are significant predictors of whether or not 1st year undergraduate students complete the questionnaire. Table 3.12 also shows that Sex, Domicile and SEC individually have p-values greater than our significance level of 0.05. Therefore they are not significant predictors of whether or not 1st year undergraduate students complete the questionnaire. These results agree with the results of Fisher's Exact Test presented above.

To investigate more complicated models to describe whether or not students complete the questionnaire, the Deviance using Generalized Likelihood Ratio Test, AIC and BIC for every

possible model shall be analysed. To compare the models the same data set must be used for every model; therefore any student with a missing demographic variable will have to be removed from the data set. SEC shall not be included in this analysis since there are 341 students whose SEC is missing and removing these would reduce the sample size too much. It has already been established above that SEC is not a significant predictor of completion.

Model	df	Deviance	AIC	BIC
Null	965	725.00	727.00	731.87
Sex	964	722.58	726.58	736.33
Age	964	720.75	724.75	734.50
Faculty	964	716.47	720.47	730.21
Domicile	962	723.88	731.88	751.37
Sex + Age	963	718.73	724.73	739.35
Sex + Faculty	963	714.69	720.69	735.31
Sex + Domicile	961	721.40	731.40	755.77
Age + Faculty	963	712.72	718.72	733.34
Age + Domicile	961	720.03	730.03	754.40
Faculty + Domicile	961	714.94	724.94	749.30
Sex + Age + Faculty	962	711.22	719.22	738.71
Sex + Age + Domicile	960	717.93	729.93	759.17
Sex + Faculty + Domicile	960	713.23	725.23	754.46
Age + Faculty + Domicile	960	711.82	723.82	753.05
Sex + Age + Faculty + Domicile	959	710.35	724.35	758.46

Table 3.13: Models for Completion at Baseline for Undergraduates

From Table 3.13 it can be seen that Age + Faculty + Domicile has the lowest deviance for a model with 3 variables, Age + Faculty has the lowest deviance for a model with 2 variables and Faculty has the lowest deviance for a model with 1 variable. Using Generalized Likelihood Ratio tests with forward selection and backwards elimination, the variation

explained by the model in Age + Faculty alone is similar to the variation explained by the other models.

When looking at AIC the table shows that the model in Age + Faculty has the lowest AIC value indicating that this would be the best model for completion. However BIC indicates that the model in Faculty alone would be the best model for completion. Stepwise Regression using AIC and Stepwise Regression using BIC were conducted. These confirmed that Age + Faculty would be best according to AIC and that Faculty would be best according to BIC.

The models in Age and Faculty and in Faculty alone have both been described as the best to describe whether or not undergraduate students complete the questionnaire. For the model in Faculty alone, Table 3.10 shows that the odds ratio for Faculty is estimated to be 2.37 with a confidence interval of (1.24, 5.00). Therefore the odds on students fully completing the questionnaire are between 1.24 and 5.00 times higher for students who are in a professional faculty than students who are in a non-professional faculty.

The model in Age and Faculty including an interaction term between the variables was also fitted. However, the interaction term was not significant with a p-value greater than our significance level of 0.05. Table 3.14 shows the fitted model in Age + Faculty using all available data; the p-value for Age has risen to 0.053 which is slightly greater than our significance level of 0.05 indicating that is not statistically significant related to whether a questionnaire is fully completed when Faculty is included in the model. The odds ratio confidence interval for Age, (0.99, 2.65), just contains 1 again indicating that Age is marginally not statistically significant.

Faculty has an odds ratio of 2.32 with a confidence interval of (1.22, 4.42) indicating that students in a profession faculty have between 1.22 and 4.22 times higher odds of completing the questionnaire than students in a non-profession faculty.

Demographic Variable	Coef	Std Error	P-value	df	Residual Deviance
Age + Faculty					
Intercept	1.434	0.228	<0.001		
Age (Under 21)	0.486	0.251	0.053		
Faculty (Profession)	0.842	0.329	0.010	966	713.86

Table 3.14: Logistic Regression of Completion for Age & Faculty at Baseline for Undergraduates

The Hosmer-Lemeshow goodness of fit test has not been calculated for the model in Faculty since it is not appropriate for a model with only 1 binary variable as described in section 2.3.4.

The Hosmer-Lemeshow goodness of fit test was performed to test if the model for the Age + Faculty is an adequate fit to the data. This produced a p-value of 0.932 which is greater than our significance level of 0.05 suggesting the model is an adequate fit.

Using the same analyses as the undergraduates, the postgraduate students shall now be looked at to establish if any of the demographic variables are a predictor of whether or not 1st year postgraduate students completed the questionnaire

Demographic Variable	Coef	Std Error	P-value	df	Residual Deviance
Sex					
Intercept	1.402	0.160	<0.001		
Sex (Males)	0.118	0.262	0.652	402	393.36
Age					
Intercept	1.308	0.159	<0.001		
Age (Under 25)	0.352	0.264	0.182	402	391.75
Faculty					
Intercept	1.356	0.159	<0.001		
Faculty (Profession)	0.239	0.264	0.366	402	392.73
Domicile					
Intercept	1.426	0.183	<0.001		
Domicile (Rest of the UK)	0.232	0.575	0.687		
Domicile (Rest of Europe)	0.047	0.381	0.901		
Domicile (Rest of the World)	0.023	0.295	0.937	392	384.22

Table 3.15: Univariate Logistic Regression of Completion by Demographic Variables at Baseline for Postgraduates

Table 3.15 also shows that all of the demographic variables individually have p-values greater than our significance level of 0.05. Therefore none of them is a significant predictor of whether or not 1st year postgraduate students complete the questionnaire.

Model	df	Deviance	AIC	BIC
Null	395	384.39	386.39	390.37
Sex	394	384.30	388.30	396.26
Age	394	382.36	386.36	394.32
Faculty	394	383.51	387.51	395.48
Domicile	392	384.22	392.22	408.14
Sex + Age	393	382.23	388.23	400.17
Sex + Faculty	393	383.46	389.46	401.41
Sex + Domicile	391	384.14	394.14	414.04
Age + Faculty	393	381.65	387.65	399.60
Age + Domicile	391	382.22	392.22	412.13
Faculty + Domicile	391	383.27	393.27	413.17
Sex + Age + Faculty	392	381.57	389.57	405.50
Sex + Age + Domicile	390	382.07	394.07	417.96
Sex + Faculty + Domicile	390	383.21	395.21	419.10
Age + Faculty + Domicile	390	381.27	393.27	417.16
Sex + Age + Faculty + Domicile	389	381.15	395.15	423.02

Table 3.16: Models for Completion at Baseline for Postgraduates

From Table 3.16 it can be seen that Age + Faculty + Domicile has the lowest deviance for a model with 3 variables, Age + Faculty has the lowest deviance for a model with 2 variables and Age has the lowest deviance for a model with 1 variable. Using Generalized Likelihood Ratio tests, the variation explained by the null model alone is similar to the variation explained by the other models.

When looking at AIC the table shows that the model of Age has the lowest AIC value indicating that this would be the best model for completion. However BIC indicates that the null model would be the best model for completion. Stepwise Regression using AIC and Stepwise Regression using BIC confirmed that Age would best according to AIC and that the null model would be best according to BIC.

The null model and the model in Age have been described as the best to describe whether or not postgraduate students complete the questionnaire. However Table 3.15 shows that Age is not a significant predictor suggesting that the null model best describes whether or not 1st year postgraduate students complete the questionnaire.

From documenting the completeness of the questionnaire, there appeared to be no pattern of missingness for item non-response. For personal attribute scale non-response, the percentages of missing personal attribute scales increased at Semester 1 and Semester 2 from Baseline; in particular, Self Efficacy and Resilience appeared to have a higher percentage of missing values consistently across each timepoint.

After formal hypothesis testing was used to examine whether or not any demographic variables appeared to be related to non-completion of each personal attribute for Undergraduates and Postgraduates for each time point the only statistically significant result was for Resilience and Sex at Baseline for Undergraduates. A higher percentage of males than females failed to complete the Resilience scale items.

When investigating the effects of demographic variables in on the completeness of the questionnaires through model building GLRT suggested that the model in Faculty alone was the best for Undergraduates. However, AIC suggested the model in Sex + Faculty + Age was the best and BIC suggested the null model as best for completion. For Postgraduates, GLRT and BIC both suggested the null model as the best for completion while AIC implied that the model in Age was the best.

It was decided that the missingness in Semester 1 and Semester 2 would not be investigated in greater detail, through model building with logistic regression, because the tests had little power due to the small number of responses and the even smaller number of missing responses at those time points.

As the completeness of the questionnaire has been documented, it would be of interest to apply methods of dealing with missing data. Complete Case analysis will be examined as a method of dealing with missing data and then compare the results of this with the results of the same analysis after multiple imputation.

Chapter 4

Complete Case Analysis

In Chapter 3 the completeness of the dataset was documented by exploring Item Non-response and questionnaire completion. In Chapter 4 the Complete Case analysis will be examined as a method of handling the missing data.

The primary aim of this study that produced these data is to investigate the proportion of 1st year students who continue at the University after the end of 1st year and their personal attributes. It is also of interest to investigate the relationship between the proportion of students who have progressed on their original degree programme and their personal attributes.

For the complete case analysis only the Undergraduate students will be investigated. This is due to the size of the Postgraduate data being too small and also because the definition of continuation and progression is complex for Postgraduates.

4.1 Continuation at Baseline

To obtain the most informative model for continuation, each possible explanatory variable (Mindset, Self Efficacy, Self Esteem, Resilience, Hope Agency, Hope Pathway, Sex, Age, Domicile, Faculty and SEC) is included in a logistic regression analysis to determine if on its own that specific explanatory is significantly related to Continuation. Then all 11 potential explanatory variables are included in the logistic regression analysis to determine if any are

significantly related to Continuation when other explanatory variables are also included in the model.

Below Table 4.1 shows the count and percentage for each of the 5 demographic variables by Continuation. From Table 4.1 it can be seen that the proportion of males that did not continue at the University of Glasgow after 1st year is higher than the proportion of females that did not continue at the University of Glasgow after 1st year, the proportion of mature students that did not continue at the University of Glasgow after 1st year is higher than the proportion of students under the age of 21 that did not continue at the University of Glasgow after 1st year. The proportion of students in a non profession faculty not continuing after 1st year is 5.39% higher than the students in a profession faculty. For domicile the proportions of students not continuing after 1st year decreases the further away from Scotland a student usually resides.

Demographic Variables		Not Continuing		Continuing	
Sex	Female	39/604	(6.46%)	565/604	(93.54%)
	Male	37/365	(10.14%)	328/365	(89.86%)
Age	Mature	15/134	(11.19%)	119/134	(88.81%)
	Under 21	61/835	(7.31%)	774/835	(92.69%)
Domicile	Scotland	56/675	(8.30%)	619/675	(91.70%)
	Rest of the UK	11/142	(7.75%)	131/142	(92.25%)
	Rest of Europe	7/113	(6.19%)	106/113	(93.81%)
	Rest of the World	2/36	(5.56%)	34/36	(94.44%)
Faculty	Non Profession	70/794	(8.82%)	724/794	(91.18%)
	Profession	6/175	(3.43%)	169/175	(96.57%)
SEC	A	23/409	(5.62%)	386/409	(94.38%)
	B	9/93	(9.68%)	84/93	(90.32%)
	C	10/129	(7.75%)	119/129	(92.25%)

Table 4.1: Continuation by Demographic Variables at Baseline

From Table 4.2 it can be seen that the median scores for each personal attribute scale are similar for 1st year students that do continue at the University of Glasgow after 1st year and 1st year students that do not continue at the University of Glasgow after 1st year, although the median is slightly higher for 1st year students that do continue for Self Esteem, Resilience and Hope Agency. Table 4.2 also shows that 1st year students that continue at the University of Glasgow after 1st year have marginally smaller interquartile ranges than students who do not continue for Mindset, Self Esteem, Resilience, Hope Total and Hope Pathway, while interquartile ranges are equal for Self Efficacy and Hope Agency.

Personal Attribute (Scale Range)	Not Continuing			Continuing		
	Median	Q1	Q3	Median	Q1	Q3
Mindset (1 to 6)	3.75	3.00	4.50	3.50	3.00	4.25
Self Efficacy (10 to 40)	31.00	29.00	34.00	31.00	29.00	34.00
Self Esteem (10 to 40)	30.00	27.00	34.00	31.00	28.00	34.00
Resilience (1 to 4)	2.90	2.70	3.13	3.00	2.80	3.20
Hope Total (8 to 32)	25.00	23.00	27.00	25.00	24.00	27.00
Hope Agency (4 to 16)	12.00	12.00	14.00	13.00	12.00	14.00
Hope Pathway (4 to 16)	12.00	11.75	13.00	12.00	12.00	13.00

Table 4.2: Continuation by Personal Attributes at Baseline

When looking at the logistic regression models for each possible explanatory variable individually, the only models to have a p-value less than our significance level of 0.05 were the models for Sex (p-value = 0.04) and for Faculty (p-value = 0.02). Therefore these explanatory variables, individually, are a significant predictor of whether or not 1st year students continue at the University of Glasgow after 1st year.

As described in Chapter 3, the Deviance using Generalized Likelihood Ratio Test, AIC and BIC for every model was analysed to determine which model best describes whether or not 1st year students continue at the University of Glasgow after 1st year. Again any student

with a missing predictor variable was removed from the dataset. As it has been established above that SEC was not a significant predictor of continuation, SEC was not included in this analysis because the sample size would be reduced too much due to the large number of students with SEC missing.

Model	df	Deviance	AIC	BIC
Null	845	463.43	465.43	470.17†
Sex	844	460.30	464.30	473.79
Age	844	460.68	464.68	474.16
Domicile	842	462.01	470.01	488.97
Faculty	844	457.72†	461.72	471.20
Mindset	844	462.87	466.87	476.35
Self Efficacy	844	463.42	467.42	476.90
Self Esteem	844	463.42	467.42	476.90
Resilience	844	462.03	466.03	475.51
Hope Agency	844	463.39	467.39	476.39
Hope Pathway	844	463.33	467.33	476.81
Sex + Age	843	457.87	463.87	478.09
Sex + Domicile	841	458.92	468.92	492.62
Sex + Faculty	843	455.19	461.19	475.41
Sex + Mindset	843	459.89	465.89	480.11
Sex + Self Efficacy	843	460.29	466.29	480.52
Sex + Self Esteem	843	460.22	466.22	480.44
Sex + Resilience	843	458.71	464.71	478.93
Sex + Hope Agency	843	460.30	466.30	480.53
Sex + Hope Pathway	843	460.27	466.27	480.52
Age + Domicile	841	458.76	468.76	492.47
Age + Faculty	843	455.29	461.29	475.51
Age + Mindset	843	460.43	466.43	480.65
Age + Self Efficacy	843	460.68	466.68	480.90
Age + Self Esteem	843	460.67	466.67	480.89
Age + Resilience	843	459.25	465.25	479.47
Age + Hope Agency	843	460.64	466.64	480.87
Age + Hope Pathway	843	460.57	466.57	480.79
Faculty + Domicile	841	455.66	465.66	489.37
Faculty + Mindset	843	457.03	463.03	477.26
Faculty + Self Efficacy	843	457.60	463.60	477.82
Faculty + Self Esteem	843	457.70	463.70	477.92
Faculty + Resilience	843	456.45	462.45	476.67
Faculty + Hope Agency	843	457.68	463.68	477.90
Faculty + Hope Pathway	843	457.48	463.48	477.70
Sex + Faculty + Age	842	452.99	460.99†	479.96
Sex + Faculty + Age + Domicile	839	450.77	464.77	497.95
Sex + Faculty + Age + Domicile + Mindset + Self Efficacy + Self Esteem + Resilience + Hope Agency + Hope Pathway	833	447.47	473.47	535.10

† The model each method indicates is the best.

Table 4.3: Some Models for Continuation at Baseline

From the above table we can see that Sex + Faculty + Age + Domicile has the lowest deviance for a model with 4 variables, Sex + Faculty + Age has the lowest deviance for a model with 3 variables, Sex + Faculty has the lowest deviance for a model with 2 variables

and Faculty has the lowest deviance for a model with 1 variable. Using Generalized Likelihood Ratio tests and comparing the best model for each number of variables (highlighted in bold), the variation explained by the model in Faculty alone is similar to the variation explained by the other models.

When looking at AIC the table shows that the model in Sex + Age + Faculty has the lowest AIC value indicating that this would be the best model for continuation. However BIC indicates that the null model would be the best model for continuation. Stepwise Regression using AIC and Stepwise Regression using BIC confirmed that Sex + Age + Faculty would be best according to AIC and that the null model would be best according to BIC.

The models in Sex + Age + Faculty and in Faculty alone have both been described as the best to describe whether or not students continue at the university after 1st year: the fitted models are displayed in Table 4.4. A model in Sex and Age and Faculty that included interaction terms among the variables was fitted. However, the interactions were not significant, with all p-values for interaction terms greater than our significance level of 0.05.

Models	Coef	Std Error	P-value	df	Residual Deviance
Faculty					
Intercept	2.336	0.125	<0.001		
Faculty (Profession)	1.002	0.434	0.021	967	525.91
Sex + Faculty + Age					
Intercept	2.193	0.307	<0.001		
Sex (Male)	-0.429	0.241	0.076		
Faculty (Profession)	0.941	0.435	0.031		
Age (Under 21)	0.400	0.307	0.192	965	520.91

Table 4.4: Logistic Regression of Continuation for Faculty and for Sex & Faculty & Age at Baseline

For the model in Faculty alone the odds ratio for Faculty is 2.72 with a confidence interval of (1.16, 6.38). Therefore the odds on students continuing at the University of Glasgow after 1st

year are between 1.16 and 6.38 times higher for students who are in a profession faculty than students who are in a non-profession faculty.

Table 4.4 shows that the p-value for Sex and Age in the additive model are greater than our significance level of 0.05 indicating that they are not statistically significant related to whether a student continues at the University of Glasgow after 1st year when Faculty is included in the model.

Faculty has an odds ratio of 2.56 with a confidence interval of (1.09, 6.01) indicating that students in a profession faculty have between 1.09 and 6.01 time higher odds of continuing at the University of Glasgow after 1st year than students in a non-profession faculty.

The Hosmer-Lemeshow test for the model in Sex + Age + Faculty produces a p-value of 0.556 which is greater than our significance level of 0.05 indicating that the model is an adequate fit to the data.

4.2 Progression at Baseline

To investigate progression at baseline, the same analysis used for continuation at baseline will be used.

Table 4.5 below shows the count and percentage for each of the 5 demographic variables by Progression. Table 4.5 shows that the proportion of males that did not progress on their original degree programme at the University of Glasgow after 1st year is higher than the proportion of females that did not progress at the University of Glasgow after 1st year, and the proportion of mature students that did not progress at the University of Glasgow after 1st year is higher than the proportion of students under the age of 21 that did not progress at the University of Glasgow after 1st year. For domicile, the proportion of students not progressing after 1st year decreases the further away from Scotland a student resides except for students that reside in the rest of the world. This is similar to the count and percentages for each of the 5 demographic variables by Continuation. Unlike continuation, the proportion of 1st year

students that did not progress on with their original degree programme in a profession faculty is higher than students in a non profession faculty.

Demographic Variables		Not Progressing		Progressing	
Sex	Female	47/604	(7.78%)	557/604	(92.22%)
	Male	47/365	(12.88%)	318/365	(87.12%)
Age	Mature	20/134	(14.93%)	114/134	(85.07%)
	Under 21	74/835	(8.86%)	761/835	(91.14%)
Domicile	Scotland	69/675	(10.22%)	606/675	(89.78%)
	Rest of the UK	13/142	(9.15%)	129/142	(90.84%)
	Rest of Europe	8/113	(7.08%)	105/113	(92.92%)
	Rest of the World	3/36	(8.33%)	33/36	(91.67%)
Faculty	Non Profession	83/794	(10.45%)	711/794	(89.55%)
	Profession	11/175	(6.29%)	164/175	(93.71%)
SEC	A	31/409	(7.58%)	378/409	(92.42%)
	B	12/93	(12.90%)	81/93	(87.10%)
	C	12/129	(9.30%)	117/129	(90.70%)

Table 4.5: Progression by Demographic Variables by at Baseline

From Table 4.6 it can be seen that the median scores for each personal attribute scale are similar for 1st year students that do progress at the University of Glasgow after 1st year and 1st year students that do not progress at the University of Glasgow after 1st year, although the median is slightly higher for 1st year students that do progress for Self Esteem, Resilience and Hope Agency. Table 4.6 also shows that the interquartile ranges are equal for Mindset, Self Efficacy, Self Esteem and Hope Agency for students that do progress at the University of Glasgow and students that do not progress at the University of Glasgow. However students that progress at the University of Glasgow have marginally smaller interquartile ranges than students who do not progress for Resilience, Hope Total and Hope Pathway.

Personal Attribute (Scale Range)	Not Progressing			Progressing		
	Median	Q1	Q3	Median	Q1	Q3
Mindset (1 to 6)	3.75	3.00	4.25	3.50	3.00	4.25
Self Efficacy (10 to 40)	31.00	29.00	34.00	31.00	29.00	34.00
Self Esteem (10 to 40)	30.00	28.00	34.00	31.00	28.00	34.00
Resilience (1 to 4)	2.90	2.70	3.18	3.00	2.80	3.20
Hope Total (8 to 32)	24.00	23.00	26.25	25.00	24.00	27.00
Hope Agency (4 to 16)	12.00	12.00	14.00	13.00	12.00	14.00
Hope Pathway (4 to 16)	12.00	11.00	13.00	12.00	12.00	13.00

Table 4.6: Progression by Personal Attributes at Baseline

For Progression, the only models to have a p-value less than our significance level of 0.05 were the models for Sex (p-value = 0.01) and for Age (p-value = 0.03), when examining logistic regression models for each possible explanatory variable individually. Therefore these explanatory variables, individually, are a significant predictor of whether or not 1st year students progressed with their original degree programme at the University of Glasgow after 1st year.

The Deviance, AIC and BIC for every model was analysed to determine which model is best to describe whether or not 1st year students progress normally after 1st year at the University of Glasgow. Again any student with a missing value was removed from the data set and SEC was not included in this analysis since it has already been established above that SEC is not a significant predictor of progression.

Model	df	Deviance	AIC	BIC
Null	845	538.53	540.53	545.27†
Sex	844	534.10	538.10	547.58
Age	844	533.98	537.98	547.46
Domicile	842	536.60	544.60	563.56
Faculty	844	536.28	540.28	549.76
Mindset	844	538.52	542.52	552.00.
Self Efficacy	844	538.44	542.44	551.92
Self Esteem	844	538.22	542.22	551.70
Resilience	844	537.15	541.15	550.63
Hope Agency	844	533.36	542.39	551.87
Hope Pathway	844	538.51	542.51	551.99
Sex + Age	843	530.01†	536.01†	550.24
Sex + Domicile	841	532.21	542.21	565.91
Sex + Faculty	843	532.31	538.31	552.53
Sex + Mindset	843	534.09	540.09	554.32
Sex + Self Efficacy	843	534.09	540.09	554.31
Sex + Self Esteem	843	534.01	540.01	554.23
Sex + Resilience	843	532.49	538.49	552.71
Sex + Hope Agency	843	534.07	540.07	554.29
Sex + Hope Pathway	843	534.03	540.03	554.25
Age + Domicile	841	531.58	541.58	565.29
Age + Faculty	843	532.00	538.00	552.22
Age + Mindset	843	533.92	539.92	554.14
Age + Self Efficacy	843	533.94	539.94	554.17
Age + Self Esteem	843	533.71	539.71	553.93
Age + Resilience	843	532.56	538.56	552.78
Age + Hope Agency	843	533.85	539.85	554.07
Age + Hope Pathway	843	533.97	539.97	554.19
Faculty + Domicile	841	533.76	543.76	567.47
Faculty + Mindset	843	536.26	542.26	556.48
Faculty + Self Efficacy	843	536.06	542.06	556.28
Faculty + Self Esteem	843	535.80	541.80	556.02
Faculty + Resilience	843	534.99	540.99	555.21
Faculty + Hope Agency	843	536.27	542.27	556.49
Faculty + Hope Pathway	843	536.28	542.28	556.50
Sex + Faculty + Age	842	528.43	536.43	555.39
Sex + Faculty + Age + Domicile	839	525.76	539.76	572.94
Sex + Faculty + Age + Domicile + Mindset + Self Efficacy + Self Esteem + Resilience + Hope Agency + Hope Pathway	833	522.81	548.81	610.44

† The model each method indicates is the best.

Table 4.7: Some Models for Progression at Baseline

Table 4.7 illustrates that Sex + Faculty + Age + Domicile has the lowest deviance for a model with 4 variables, Sex + Faculty + Age has the lowest deviance for a model with 3 variables, Sex + Age has the lowest deviance for a model with 2 variables and Age has the lowest deviance for a model with 1 variable. Using the Generalized Likelihood Ratio test, the variation explained by the model of Sex + Age alone is similar to the variation explained by the other models.

When looking at AIC the table shows that the model of Sex + Age has the lowest AIC value indicating that this would be the best model for progression. However BIC indicates that the null model would be the best model for progression. Stepwise Regression using AIC and Stepwise Regression using BIC confirmed that Sex + Age would best according to AIC and that the null model would be best according to BIC.

Model	Coef	Std Error	P-value	df	Residual Deviance
Sex + Age					
Intercept	2.008	0.272	<0.001		
Sex (Male)	-0.534	0.219	0.015		
Age (Under 21)	0.544	0.273	0.046	966	606.94

Table 4.8: Logistic Regression of Progression for Sex & Age at Baseline

Table 4.8 shows the fitted additive model for Sex and Age. The interaction model for Sex and Age was also fitted but the interaction term was not statistically significant. The p-values for both Age and Sex in the additive model are less than 0.05, therefore each of them has a significant effect on the probability of 1st year students progressing with their original degree programme at the University of Glasgow after 1st year in addition to the other. The odds ratio for Sex is 0.59 with a confidence interval of (0.38, 0.90), signifying that for any given Age group the odds of a female student progressing with their original degree programme at the University of Glasgow after 1st year are between 1.11 and 2.63 times higher than a male student. The odds ratio for Age is 1.72 with a confidence interval of (1.01, 2.94). Therefore for any given gender the odds of students who progress are between 1.01 and 2.94 times higher for students under 21 than students who are mature.

The Hosmer-Lemeshow test for the model in Sex + Age produces a p-value of 0.563 which is greater than our significance level of 0.05 indicating that the model is an adequate fit to the data.

There is uncertainty about which model is best for both Continuation and Progression at baseline, so potentially it is useful to impute the data that are missing (12.7% of all the possible responses).

4.3 Continuation at Baseline and Semester 1/Semester 2 for Difference in Personal Attribute Scores

The relationship of Continuation with the difference in each personal attribute score at Baseline and Semester 1/Semester 2 shall now be examined. The Difference in score was calculated by subtracting the Baseline score for each personal attribute from the same student's Semester 1/Semester 2 score, with a negative value therefore signifying a decrease in score and a positive value signifying an increase in score.

Continuation shall be investigated using the same methods as in section 4.1, replacing the baseline Personal Attribute score with the difference in Personal Attribute score (denoted as δ). The individual univariate personal attribute models will be explored with and without a binary variable indicating the Semester in which the second response was obtained. It was decided that the paired differences in Semester 1 and the paired differences in Semester 2, which were all obtained from different students, would be combined as the datasets were too small to model individually.

Personal Attribute (Scale Range)	Not Continuing			Continuing		
	Median	Q1	Q3	Median	Q1	Q3
δMindset (-6 to 6)	-0.25	-0.75	0.5	-0.25	-0.75	0.25
δSelf Efficacy (-40 to 40)	-1.50	-3.25	1.25	0.00	-2.00	1.00
δSelf Esteem (-40 to 40)	-2.00	-4.00	0.00	0.00	-3.00	1.00
δResilience (-4 to 4)	0.10	-0.10	0.10	-0.10	-0.20	0.10
δHope Total (-32 to 32)	-0.50	-2.00	1.25	0.00	-2.00	1.00
δHope Agency (-16 to 16)	0.00	-2.25	1.00	0.00	-1.00	0.00
δHope Pathway (-16 to 16)	0.00	-1.25	0.50	0.00	-1.00	1.00

Table 4.9: Continuation by Difference in Personal Attributes

With the exception of Mindset and Resilience, Table 4.9 illustrates that the 1st year students who continue at the University of Glasgow after 1st year have a median of 0, indicating that there is no systematic difference in their Personal Attribute score. For the 1st year students who do not continue at the University of Glasgow after 1st year the median difference in Mindset, Self Efficacy, Self Esteem and Hope Total is negative indicating that the Personal Attribute scores have decreased. From Table 4.9 it can also be seen that students who continue at the University of Glasgow have a marginally smaller interquartile range for the difference in score for Mindset, Self Efficacy, Hope Total and Hope Agency than those 1st year students who do not continue. The opposite occurs for the difference in Resilience and the difference in Hope Pathway, where students who progress at the University of Glasgow have a marginally larger interquartile range.

All p-values for the univariate logistic regressions models, for each possible explanatory variable, were greater than our significance level of 0.05. Therefore each explanatory variable individually is not a significant predictor of whether or not 1st year students continue at the University of Glasgow after 1st year. From the univariate logistic regression models, it was also established that the semester indicator variable was not statistically significant and it will be removed for the model building.

Using the same methods as in section 4.1, the Deviance using Generalized Likelihood Ratio Test, AIC and BIC for every model was analysed to determine which model best describes whether or not 1st year students continue at the University of Glasgow after 1st year. As before SEC is not included in this analysis as the sample size would be reduced too much and it has been established that SEC is not a significant predictor of continuation.

Due to the large number of models that were fitted Table 4.10 contains a selection of models with potential interesting Deviance, AIC and BIC values. There was no model that contained more than 1 of the difference in personal attribute score that was approximately the best model.

Model	df	Deviance	AIC	BIC
Null	264	109.59†	111.59	115.17†
Sex	263	107.64	111.64	118.80
Age	263	109.45	113.45	120.61
Domicile	261	104.72	112.72	127.04
Faculty	263	107.51	111.51	118.67
δMindset	263	109.52	113.52	120.68
δSelf Efficacy	263	108.90	112.90	120.06
δSelf Esteem	263	107.09	111.09	118.25
δResilience	263	109.43	113.43	120.59
δHope Agency	263	108.40	112.40	119.56
δHope Pathway	263	109.57	113.57	120.73
Sex + Age	262	107.60	113.60	124.34
Sex + Domicile	260	102.64	112.64	130.54
Sex + Faculty	262	106.01	112.01	122.75
Sex + δMindset	262	107.34	113.34	124.08
Sex + δSelf Efficacy	262	106.79	112.79	123.53
Sex + δSelf Esteem	262	105.07	111.07†	121.80
Sex + δResilience	262	107.57	113.57	124.31
Sex + δHope Agency	262	106.15	112.15	122.89
Sex + δHope Pathway	262	107.61	113.61	124.35
Age + Domicile	260	104.46	114.46	132.36
Age + Faculty	262	107.42	113.42	124.16
Age + δMindset	262	109.41	115.41	126.14
Age + δSelf Efficacy	262	108.71	114.71	125.45
Age + δSelf Esteem	262	106.95	112.95	123.69
Age + δResilience	262	109.32	115.32	126.06
Age + δHope Agency	262	108.31	114.31	125.05
Age + δHope Pathway	262	109.44	115.44	126.48
Faculty + Domicile	260	102.88	112.88	130.78
Faculty + δMindset	262	107.41	113.41	124.15
Faculty + δSelf Efficacy	262	106.74	112.74	123.48
Faculty + δSelf Esteem	262	105.07	111.07	121.81
Faculty + δResilience	262	107.36	113.36	124.10
Faculty + δHope Agency	262	106.45	112.45	123.19
Faculty + δHope Pathway	262	107.50	113.50	124.24
Sex + Faculty + Age	261	105.97	113.97	128.29
Sex + Faculty + δSelf Esteem	261	103.52	111.52	125.84
Sex + Faculty + Age + Domicile	258	101.22	115.22	140.27
Sex + Faculty + Age + δSelf Esteem	260	103.44	113.44	131.34
Sex + Faculty + Age + Domicile + δMindset + δSelf Efficacy + δSelf Esteem + δResilience + δHope Agency + δHope Pathway	252	97.04	123.04	169.58

† The model each method indicates is the best.

Table 4.10: Some Models for Continuation at Baseline and Semester1/2

Using Generalized Likelihood Ratio tests with forward selection and backwards elimination, the variation explained by the null model is similar to the variation explained by the other models suggesting this is best model to describe continuation.

When looking at AIC the table shows that the model of Sex + δ Self Esteem has the lowest AIC value indicating that this would be the best model for continuation. However BIC indicates that the null model would be the best model for continuation. Stepwise Regression using AIC and Stepwise Regression using BIC confirmed these results.

Model	Coef	Std Error	P-value	df	Residual Deviance
Sex + δSelf Esteem					
Intercept	3.455	0.397	<0.001		
Sex	-0.685	0.534	0.200		
δ Self Esteem	0.133	0.083	0.109	321	117.29

Table 4.11: Logistic Regression of Continuation for Sex & Difference in Self Esteem at Baseline and Semester1/2

A model in Sex and δ Self Esteem that an included interaction term was fitted, however the interaction term was not statistically significant. Shown in Table 4.11 is the additive model for Sex and δ Self Esteem. The p-values for Sex and δ Self Esteem are greater than our significance level of 0.05 indicating that they are not statistically significant related to whether a student continues at the University of Glasgow after 1st year.

4.4 Progression at Baseline and Semester 1/ Semester 2 for Difference in Personal Attribute Scores

The relationship of Progression and the difference in each personal attribute score at Baseline and Semester 1/Semester 2 shall also be examined. Progression shall be investigated using the same methods in 4.3.

Personal Attribute (Scale Range)	Not Progressing			Progressing		
	Median	Q1	Q3	Median	Q1	Q3
δMindset (-6 to 6)	-0.25	-0.75	0.50	-0.25	-0.75	0.25
δSelf Efficacy (-40 to 40)	-1.00	-3.00	1.00	0.00	-2.00	1.00
δSelf Esteem (-40 to 40)	-1.50	-5.75	0.00	0.00	-3.00	1.00
δResilience (-4 to 4)	0.05	-0.10	0.10	-0.10	-0.20	0.10
δHope Total (-32 to 32)	0.00	-2.00	1.00	0.00	-2.00	1.00
δHope Agency (-16 to 16)	0.00	-2.50	1.00	0.00	-1.00	0.00
δHope Pathway (-16 to 16)	0.00	-0.50	0.00	0.00	-1.00	1.00

Table 4.12: Difference in Personal Attributes by Progression

With the exception of Mindset and Resilience, Table 4.12 illustrates that the 1st year students who progress on to the next year of their original degree programme at the University of Glasgow after 1st year have a median of 0, indicating that there is no difference in their Personal Attribute score. For the 1st year students who do not progress at the University of Glasgow the median for difference in Self Efficacy and Self Esteem is negative indicating that the Personal Attribute scores have decreased. From Table 4.12 it can also be seen that students who progress at the University of Glasgow have a marginally smaller interquartile range for the difference in score for Mindset, Self Efficacy, Self Esteem and Hope Agency than those 1st year students who do not progress. The opposite occurs for the difference in Resilience and the difference in Hope Pathway, where students who progress at the University of Glasgow have a marginally larger interquartile range.

When looking at the univariate logistic regression models for each possible explanatory variable, the only model to have a p-value less than our significance level of 0.05 was the model including the difference in Self Esteem. Therefore the difference in Self Esteem is a significant predictor of whether or not 1st year students progress on their original degree program after 1st year.

Using the same methods as in section 4.2, the Deviance using Generalized Likelihood Ratio Test, AIC and BIC for every model was analysed to determine which model best describes whether or not 1st year students progress at the University of Glasgow after 1st year. As before SEC is not included in this analysis as the sample size would be reduced too much and it has been established that SEC is not a significant predictor of progression.

Due to the large number of models that were fitted Table 4.13 contains a selection of models with potential interesting Deviance, AIC and BIC values. There was no model that contained more than 1 of the difference in personal attribute score that was approximately the best model.

Model	df	Deviance	AIC	BIC
Null	264	146.77	148.77	152.61†
Sex	263	145.69	149.69	157.38
Age	263	146.23	150.23	157.92
Domicile	261	141.44	149.44	164.82
Faculty	263	146.64	150.64	158.33
δMindset	263	145.53	149.53	157.21
δSelf Efficacy	263	146.25	150.25	157.93
δSelf Esteem	263	141.09†	145.09†	152.77
δResilience	263	146.61	150.61	158.30
δHope Agency	263	145.81	149.81	157.50
δHope Pathway	263	146.76	150.76	158.44
Sex + Age	262	145.32	151.32	162.85
Sex + Domicile	260	140.23	150.23	169.44
Sex + Faculty	262	145.43	151.43	162.96
Sex + δMindset	262	143.82	149.82	161.35
Sex + δSelf Efficacy	262	145.08	151.08	162.61
Sex + δSelf Esteem	262	139.90	145.90	157.43
Sex + δResilience	262	145.60	151.60	163.13
Sex + δHope Agency	262	144.54	150.54	162.07
Sex + δHope Pathway	262	145.68	151.68	163.21
Age + Domicile	260	140.84	150.84	170.06
Age + Faculty	262	146.08	152.08	163.61
Age + δMindset	262	145.08	151.08	162.61
Age + δSelf Efficacy	262	145.62	151.62	163.15
Age + δSelf Esteem	262	140.53	146.53	158.06
Age + δResilience	262	146.11	152.11	163.64
Age + δHope Agency	262	145.34	151.34	162.97
Age + δHope Pathway	262	146.23	152.23	163.76
Faculty + Domicile	260	141.44	151.44	170.66
Faculty + δMindset	262	145.44	151.44	162.97
Faculty + δSelf Efficacy	262	146.14	152.14	163.67
Faculty + δSelf Esteem	262	140.92	146.92	158.45
Faculty + δResilience	262	146.49	152.49	164.02
Faculty + δHope Agency	262	145.65	151.65	163.19
Faculty + δHope Pathway	262	146.63	152.63	164.16
Domicile + δSelf Esteem	260	136.50	146.50	165.72
Sex + Faculty + Age	261	145.05	153.05	168.43
Sex + Faculty + δSelf Esteem	261	139.56	147.56	162.93
Sex + Faculty + Age + Domicile	258	139.75	153.75	180.66
Sex + Faculty + Domicile + δSelf Esteem	258	134.86	148.86	175.76
Sex + Faculty + Age + Domicile + δMindset + δSelf Efficacy + δSelf Esteem + δResilience + δHope Agency + δHope Pathway	252	130.91	156.91	206.88

† The model each method indicates is the best.

Table 4.13: Some Models for Progression at Baseline and Semester1/2

Generalized Likelihood Ratio tests were used to compare all the models. The variation explained by the model of δ Self Esteem alone is similar to the variation explained by the other models recommending this as the best model.

When looking at AIC the table shows that the model in δ Self Esteem has the lowest AIC value indicating that this would be the best model for progression. However BIC indicates the null model would be the best model for progression. Stepwise Regression using AIC and Stepwise Regression using BIC confirmed this.

Model	Coef	Std Error	P-value	df	Residual Deviance
δSelf Esteem					
Intercept	2.851	0.266	<0.001		
δ Self Esteem	0.169	0.069	0.014	322	154.67

Table 4.14: Logistic Regression of Progression for Difference in Self Esteem at Baseline and Semester1/2

Table 4.14 shows that the p-value for the difference in Self Esteem is less than our significant level of 0.05, therefore the difference in Self Esteem is a significant predictor of whether or not 1st year students progress on their original degree program after 1st year. The coefficient value for the difference in Self Esteem is positive indicating that the odds of a student progressing on to the next year of their original degree program after 1st year increases as the difference in Self Esteem increases. The odds ratio for the difference in Self Esteem is 1.84 with a confidence interval of (1.04, 1.36), signifying that for a 1 unit increase in the difference in Self Esteem the odds of student progressing at the University of Glasgow after 1st year are between 1.04 and 1.36 higher. Figure 4.1 below shows the probability of Progressing by the difference in Self Esteem score. Highlighted in bold is the difference in Self Esteem score from -12 to 12 as this is maximum and minimum difference in Self Esteem score recorded shown in Table 4.12. Figure 4.1 shows that as the difference in Self Esteem increases the probability of progressing increases.

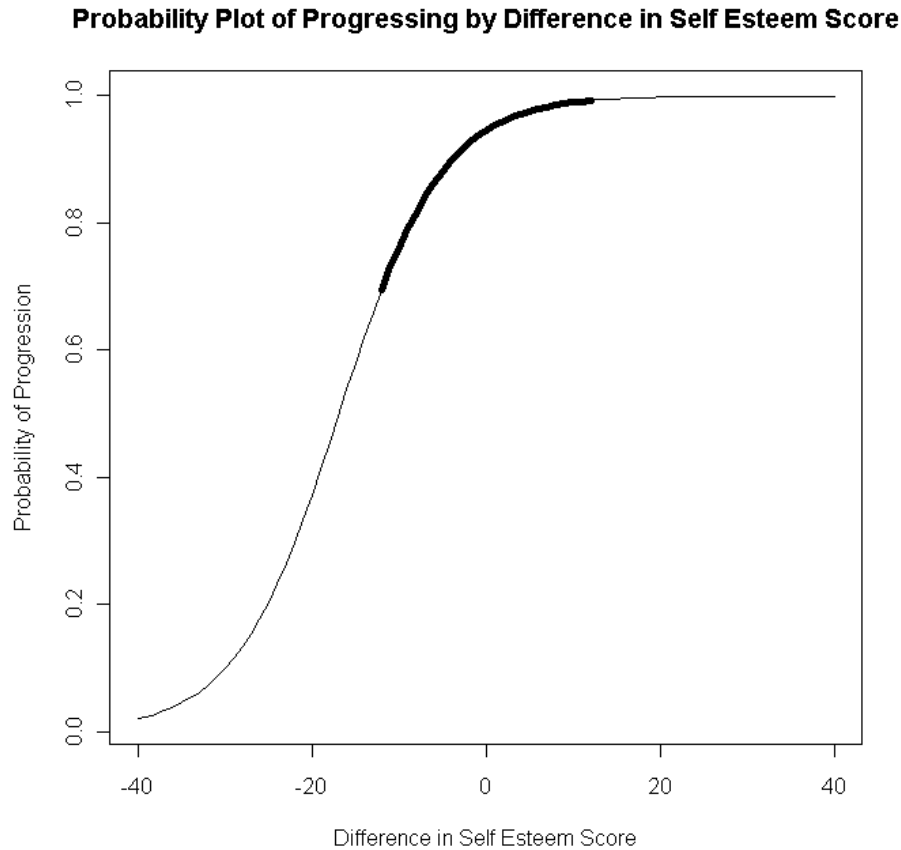


Figure 4.1: Probability of Progression by Difference in Self Esteem Score

The Hosmer-Lemeshow goodness of fit test was performed to test if the model for the difference in Self Esteem is an adequate fit to the data. This produced a p-value of 0.162 which is greater than our significance level of 0.05 suggesting the model is an adequate fit.

Throughout the Complete Case analysis there was uncertainty about which model is best for both Continuation and Progression at Baseline and for both Continuation and Progression at Baseline and Semester 1/Semester 2. For Continuation at Baseline there was no agreement between the three model building methods as GLRT suggested the model in Faculty alone, AIC suggested the model of Sex + Faculty + Age and BIC suggested the null model. For Progression at Baseline both GLRT and AIC suggested the model of Sex + Age while BIC suggested the null model. For Continuation at Baseline and Semester 1/Semester 2 the null model was suggested by both GLRT and BIC whereas AIC suggested the additive model for

Sex and difference in Self Esteem. GLRT and AIC both suggested that model in difference in Self Esteem as the best for Progression at Baseline and Semester 1/Semester 2. However BIC indicated that the null model would be best. It may be potentially useful to try using imputation to create a larger sample that can be used for the same analysis described in Chapter 4.

Chapter 5

Multiple Imputation

Multiple imputation is a technique that involves filling-in missing data repeatedly to create a set of $D \geq 2$ complete datasets (where $D = 10$ for this thesis). It is of interest to compare the results from the Complete Case analyses in Chapter 4 when more complex missing data techniques are implemented to impute missing personal attribute values.

For multiple imputation, it was decided that imputing at Baseline only would be explored first, before attempting to impute all of the Baseline, Semester 1 and Semester 2 data. When imputing at Baseline, missing overall scale values were imputed to begin with as this was the simplest option. It was then decided to attempt to obtain more accurate results by imputing individual item values.

When moving on to imputing the Baseline, Semester 1 and Semester 2 personal attribute data, the imputations were split into two steps. Semester 1 and Semester 2 item values were combined with a semester indicator variable in the same manner described in Section 4.3. The first step was to try item-level imputation but only for students who had attempted a follow up questionnaire. The second step was to impute scale values for Semester 1/Semester 2 for the students that only attempted a Baseline questionnaire starting from one of the imputed datasets from the first step combined with an imputed dataset where items had been imputed at Baseline only.

For all multiple imputations conducted in this thesis no missing demographic variables will be imputed, only the personal attribute values will be imputed. Instead any missing

demographic variable will not be treated as missing and be classed as an “unknown” category.

As in Chapter 4 only the Undergraduate students will be investigated and imputed.

5.1 Imputing Scale Level Data at Baseline

The numbers of missing values for each scale at each time point are listed in Table 3.7. As there are missing values for individual items and also for entire personal attribute scales, it needs to be decided whether it is best to impute at the scale level or at the question level.

To start with, the missing scale values shall be imputed for all students that participated in the study, regardless of if they have non-missing item data or not, as this is the simplest and cheapest option. As Hope Agency plus Hope Pathway equals Hope Total, Hope Total will not be included in the imputation. As well as the 6 scales, the imputation model will include Sex, Age, Faculty, Domicile and SEC as predictors. No limitations will be set for the imputed values, although there are well-defined minimum and maximum values for all the scales. This is because it is not possible within the **mi package** and to keep the imputation as simple as possible.

10 imputations were conducted with each having a maximum of 50 iterations to establish if the imputation has converged.

To establish how well this imputation worked, the plausibility of the imputed scale values was checked. The imputed scale values were classed as plausible if the imputed value lay between the minimum and maximum possible values of the scale.

From Table 5.1 it can be seen that very few imputed scale values are classed as not being plausible which is encouraging considering that no limitations were imposed on the imputed values. For example, 9 missing values of Mindset had to be imputed; in 8 of the 10 imputations, all 9 imputed values were plausible and in the other 2 imputations, just one of the imputed values fell outside the range of the Mindset scale.

Imputed Data Set	Mindset		Self Efficacy		Self Esteem		Resilience		Hope Agency		Hope Pathway	
	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
1	0	9	0	41	1	24	0	47	0	18	0	15
2	0	9	0	41	1	24	0	47	1	17	0	15
3	0	9	1	40	0	25	0	47	1	17	0	15
4	1	8	0	41	1	24	0	47	0	18	0	15
5	0	9	0	41	0	25	0	47	0	18	0	15
6	1	8	0	41	0	25	0	47	0	18	0	15
7	0	9	0	41	0	25	0	47	0	18	0	15
8	0	9	0	41	0	25	0	47	0	18	0	15
9	0	9	0	41	0	25	0	47	0	18	0	15
10	0	9	0	41	0	25	0	47	1	17	0	15

Table 5.1: Plausibility for Scale Level Data at Baseline

To assess the quality of each imputed scale value, the imputed scale value was compared to a range obtained from the non-missing question responses by the given student. If a student has not missed out every question within a scale, then the total of that student's given responses can be calculated. Using this total, a minimum and maximum range can be calculated as follows:

$$\text{Minimum} = \text{Total} + (\text{Lowest Response Value}) \times (\text{No. of Missing Questions})$$

$$\text{Maximum} = \text{Total} + (\text{Highest Response Value}) \times (\text{No. of Missing Questions})$$

Table 5.2 below gives an example of calculating the consistent range for Self Esteem, where each completed item is scored between 1 and 4.

<i>Q1</i>	<i>Q2</i>	<i>Q3</i>	<i>Q4</i>	<i>Q5</i>	<i>Q6</i>	<i>Q7</i>	<i>Q8</i>	<i>Q9</i>	<i>Q10</i>	<i>Total</i>	<i>Min</i>	<i>Max</i>	<i>Imputed Value</i>	<i>Consistent</i>
2	2	3	?	3	4	4	2	2	4	26	27	30	31	No
2	3	3	3	?	2	2	?	3	2	20	22	28	26	Yes

Where for row 1 $\text{Min} = 26 + 1 \times (1) = 27$ and $\text{Max} = 26 + 4 \times (1) = 30$
and for row 2 $\text{Min} = 20 + 1 \times (2) = 22$ and $\text{Max} = 20 + 4 \times (2) = 28$

Table 5.2: Example of Consistency

The imputed scale values are classed as consistent if the imputed value lies within the range of the minimum and maximum score.

Imputed Data Set	Mindset		Self Efficacy		Self Esteem		Resilience		Hope Agency		Hope Pathway	
	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
1	7	2	27	14	18	7	32	15	10	8	5	10
2	6	3	32	9	18	7	26	21	6	12	4	11
3	5	4	27	14	16	9	39	8	6	12	5	10
4	5	4	23	18	18	7	32	15	10	8	4	11
5	5	4	27	14	18	7	31	16	9	9	3	12
6	5	4	24	17	19	6	31	16	3	15	4	11
7	4	5	29	12	18	7	35	12	7	11	2	13
8	3	6	22	19	17	8	34	13	9	9	1	14
9	4	5	26	15	18	7	33	14	11	7	3	12
10	7	2	25	16	17	8	35	12	7	11	5	10

Table 5.3: Consistency for Scale Level Data at Baseline

Imputed Data Set	Mindset		Self Efficacy		Self Esteem		Resilience		Hope Agency		Hope Pathway	
	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High
1	3	4	11	16	4	14	12	20	4	6	3	2
2	2	4	13	19	6	12	12	14	1	5	3	1
3	2	3	12	15	9	7	16	23	12	6	4	1
4	1	4	12	11	6	12	12	20	0	10	2	2
5	2	3	10	17	8	10	11	20	4	5	1	2
6	2	3	11	13	12	7	12	19	0	3	2	2
7	0	4	9	20	7	11	15	20	2	5	1	1
8	1	2	8	14	5	12	10	24	1	8	0	1
9	1	3	13	13	8	10	17	16	5	6	2	1
10	2	5	14	11	9	8	16	19	3	4	2	3

Table 5.4: Direction of Non Consistent Values for Scale Level Data at Baseline

From Table 5.3 it can be seen that over half the imputed values are not consistent for each scale apart from Hope Pathway. Further investigation with Table 5.4 shows that over half of the non-consistent values are being over estimated.

From this it can be established that this simple form of imputation at the scale level is not accurate enough to predict the missing values. Instead the imputation at the scale level should consider including the information about the question values or perhaps it would be better to impute at the item level instead.

5.2 Imputing Item Level Data at Baseline

As imputing at the scale level was not an accurate enough method to predict the missing values, imputation at the item level shall be looked at. It was decided that the items would be imputed for each personal attribute scale separately when imputing at the item level. For Hope it was decided not to split the items into Hope Agency and Hope Pathway and instead have all 12 Hope items grouped together. For each of the imputation models, predictors included Sex, Age, Faculty, Domicile, SEC and the items for the given scale. For each of the imputations, the imputed item values were treated as ordered categorical variables.

10 imputations for each personal attribute scale were conducted with each having a maximum of 1000 iterations to establish if the imputation has converged. Once the 10 imputed data sets for each personal attribute scale had been obtained the imputed values were all checked and found to be plausible. To establish full imputed data sets, one of the imputed personal attribute scales data sets for each scale was randomly selected, without replacement, to be combined into a full data set. This resulted in 10 complete imputed data sets and the personal attribute scale scores were calculated in each. The same analysis used in the complete case analysis was applied to the 10 imputed data sets separately.

5.2.1 Continuation for Imputed Item Level Data at Baseline

To begin with, the univariate analysis using each individual personal attribute scale shall be examined for all 10 imputed data sets. The univariate analysis for the 5 demographic variables is the same for all of the imputed datasets because no demographic information had to be imputed.

Below (Tables 5.5 and 5.6) are parameter estimates from the univariate logistic regression analysis in each individual personal attribute, for all 10 imputed datasets separately then combined (using equation 2.3, 2.4, 2.5 & 2.6 in Chapter 2).

From Table 5.5 it can be seen that for each of the personal attribute scales, there is substantial consistency for the slope estimate across all imputations. All p-values for the models

displayed in Table 5.5, for each possible explanatory variable, were greater than our significance level of 0.05. Therefore each explanatory variable individually is not a significant predictor of whether or not 1st year students continue at the University of Glasgow after 1st year.

Table 5.6 also shows that there is substantial consistency for the slope estimate as the between-imputation variability is very small for each of the personal attribute scales. It can also be seen that the combined slope estimates for the imputed datasets are very similar to the complete case slope estimate for each of the personal attributes and that the within-imputation variance is small except for Resilience. For Resilience, the within-imputation variance is quite large indicating that there is less precision in the slope parameter estimate. The within-imputation variance appears to quite large due to the standard error of the slope being quite large across all the imputed datasets. However it is unclear why the standard error of the slope is quite large across all the imputed datasets.

		1		2		3		4		5		6		7		8		9		10	
Personal Attribute		Coef	Std Error	Coef	Std Error	Coef	Std Error	Coef	Std Error	Coef	Std Error	Coef	Std Error	Coef	Std Error	Coef	Std Error	Coef	Std Error	Coef	Std Error
Mindset	Intercept	2.836	0.484	2.855	0.485	2.887	0.485	2.893	0.486	2.885	0.485	2.860	0.485	2.854	0.484	2.890	0.486	2.866	0.485	2.860	0.485
	Slope	-0.102	0.127	-0.107	0.127	-0.115	0.127	-0.117	0.127	-0.115	0.127	-0.108	0.127	-0.106	0.127	-0.116	0.127	-0.110	0.127	-0.108	0.127
Self Efficacy	Intercept	2.420	1.114	2.565	1.123	2.529	1.119	2.501	1.114	2.402	1.115	2.505	1.118	2.541	1.119	2.426	1.114	2.462	1.114	2.438	1.115
	Slope	0.001	0.035	-0.003	0.036	-0.002	0.035	-0.001	0.035	0.002	0.035	-0.001	0.035	-0.002	0.035	0.001	0.035	0.000	0.035	0.001	0.035
Self Esteem	Intercept	2.432	0.789	2.432	0.789	2.431	0.789	2.435	0.788	2.409	0.790	2.421	0.789	2.428	0.789	2.427	0.789	2.429	0.788	2.424	0.790
	Slope	0.001	0.026	0.001	0.026	0.001	0.026	0.001	0.026	0.002	0.026	0.001	0.026	0.001	0.026	0.001	0.026	0.001	0.026	0.001	0.026
Resilience	Intercept	1.416	1.157	1.483	1.157	1.430	1.160	1.399	1.157	1.342	1.159	1.449	1.157	1.387	1.157	1.424	1.159	1.443	1.156	1.508	1.157
	Slope	0.353	0.39	0.330	0.389	0.348	0.390	0.359	0.390	0.378	0.390	0.342	0.389	0.363	0.390	0.350	0.390	0.344	0.389	0.322	0.389
Hope Agency	Intercept	1.943	0.855	1.984	0.860	2.055	0.862	2.051	0.862	2.040	0.863	2.055	0.863	2.016	0.856	1.994	0.859	2.032	0.862	2.061	0.862
	Slope	0.041	0.067	0.038	0.067	0.032	0.067	0.032	0.067	0.033	0.067	0.032	0.067	0.035	0.067	0.037	0.067	0.034	0.067	0.032	0.067
Hope Pathway	Intercept	2.803	1.005	2.796	1.005	2.810	1.006	2.816	1.007	2.789	1.004	2.802	1.003	2.803	1.006	2.808	1.004	2.824	1.008	2.814	1.004
	Slope	-0.028	0.082	-0.027	0.082	-0.028	0.082	-0.029	0.082	-0.027	0.081	-0.028	0.081	-0.028	0.082	-0.028	0.081	-0.029	0.082	-0.029	0.081

Table 5.5: Univariate Logistic Regression of Continuation by Personal Attribute Scales at Baseline

Personal Attribute	Complete Case Slope Estimate	Combined Slope Estimate $\bar{\theta}_D$	Within-imputation Variance \bar{W}_D	Between-imputation Variance \bar{B}_D	Total Variance \bar{T}_D	95% Confidence Interval
Mindset	-0.114	-0.110	0.016	2.56e-5	0.016	(-0.360, 0.139)
Self Efficacy	0.001	0.000	0.001	2.71e-6	0.001	(-0.069, 0.068)
Self Esteem	0.008	0.001	0.001	1.00e-7	0.001	(-0.050, 0.052)
Resilience	0.611	0.349	0.152	2.58e-4	0.152	(-0.415, 1.113)
Hope Agency	0.030	0.035	0.004	9.82e-6	0.004	(-0.097, 0.166)
Hope Pathway	-0.032	-0.028	0.007	5.44e-7	0.007	(-0.188, 0.131)

Table 5.6: Combined Estimates and Sampling Variability of Continuation by Personal Attribute Scales at Baseline

The same model building analysis used in Chapter 4 was used on all of the imputed datasets, with the personal attribute scales and the demographic variables included as possible predictor variables. All of the imputed datasets suggested that, when using Generalized Likelihood Ratio Tests with forward selection and backwards elimination, the model in Faculty alone is the best model to describe continuation. BIC values also suggested that the model in Faculty alone is the best model to describe continuation for all of the imputed dataset. However when looking at AIC, all of the imputed datasets suggest that the model in Sex + Faculty best describes continuation. These preferred models contain none of the personal attribute scales.

The models in Sex + Faculty and in Faculty alone have both been described as the best to describe whether or not students continue at the university after 1st year. The fitted model for Sex + Faculty is displayed in Table 5.7 and the fitted model for Faculty alone is displayed in Table 4.4 as this was suggested by GLRT in the Complete Case analysis as the best model. (These fitted models are the same for all of the imputed datasets because no demographic variables had to be imputed.) A model in Sex and Faculty that included the interaction term was fitted; however the interaction was not significant as the p-value for the interaction term was greater than our significance level of 0.05.

Models	Coef	Std Error	P-value	df	Residual Deviance	AIC	BIC
Sex + Faculty							
Intercept	2.533	0.171	<0.001				
Sex (Male)	-0.446	0.241	0.064				
Faculty (Profession)	0.955	0.435	0.028	966	521.96	527.96	543.13

Table 5.7: Logistic Regression of Continuation for Sex & Faculty at Baseline

For the model in Faculty alone, as described in the Complete Case analysis, the odds ratio for Faculty is 2.72 with a confidence interval of (1.16, 6.38). Therefore the odds on students continuing at the University of Glasgow after 1st year are between 1.16 and 6.38 times higher for students who are in a profession faculty than students who are in a non-profession faculty.

Table 5.7 shows that the p-value for Sex in the additive model is greater than our significance level of 0.05 indicating that it is not statistically significant related to whether a student continues at the University of Glasgow after 1st year when Faculty is included in the model.

Faculty has an odds ratio of 2.60 with a confidence interval of (1.11, 6.10) indicating that students in a profession faculty have between 1.11 and 6.10 time higher odds of continuing at the University of Glasgow after 1st year than students in a non-profession faculty (after correction for Sex).

When comparing the model building results above to the complete case analysis in section 4.1 it can be seen that when using Generalized Likelihood Ratio Tests with forward selection and backwards elimination both the complete case and the full dataset chose the model in Faculty. However for BIC the full dataset now suggested the model in Faculty instead of the null model that was chosen in the complete case analysis and AIC has suggested Sex + Faculty instead of Sex + Age + Faculty. There is more agreement among the 3 methods on which model is best; this might be due to more data being available for the model building process.

5.2.2 Progression for Imputed Item level Data at Baseline

The same analysis used for continuation in section 5.2.1 will be repeated for progression. In the univariate analysis using the demographic variables, the only models to have a p-value less than our significance level of 0.05 included Sex or Age, indicating that these explanatory variables are individually significant predictors of progression.

Table 5.8 shows the results of the univariate logistic regression analysis in each individual personal attribute and Table 5.9 shows the combined estimates.

For the univariate logistic regressions there is substantial consistency for the slope estimate across all imputations for each of the personal attribute scales. In all of the imputed dataset, each of the explanatory variables individually is not a significant predictor of whether or not

1st year students progress to the next year of their original degree programme at the University of Glasgow after 1st year.

There is substantial consistency for the slope estimate as the between-imputation variability is very small for each of the personal attribute scales. Also the combined slope estimates for the imputed datasets are very similar to the complete case slope estimate for each of the personal attributes and the within-imputation variance is small except for Resilience. For Resilience, the within-imputation variance is quite large indicating that there is less precision in the slope parameter estimate. This is similar to the univariate logistic regression analysis for continuation.

		1		2		3		4		5		6		7		8		9		10	
Personal Attribute		Coef	Std Error	Coef	Std Error	Coef	Std Error	Coef	Std Error	Coef	Std Error	Coef	Std Error	Coef	Std Error	Coef	Std Error	Coef	Std Error	Coef	Std Error
Mindset	Intercept	2.200	0.430	2.216	0.430	2.241	0.430	2.246	0.431	2.239	0.430	2.218	0.430	2.216	0.430	2.242	0.431	2.224	0.430	2.220	0.430
	Slope	0.008	0.115	0.004	0.115	-0.003	0.115	-0.004	0.115	-0.002	0.115	0.003	0.115	0.004	0.115	-0.003	0.115	0.002	0.115	0.003	0.115
Self Efficacy	Intercept	2.191	1.012	2.341	1.020	2.246	1.015	2.258	1.012	2.172	1.013	2.261	1.015	2.323	1.017	2.197	1.012	2.226	1.012	2.233	1.014
	Slope	0.001	0.032	-0.003	0.032	0.000	0.032	-0.001	0.032	0.002	0.032	-0.001	0.032	-0.003	0.032	0.001	0.032	0.000	0.032	0.000	0.032
Self Esteem	Intercept	2.569	0.728	2.569	0.728	2.567	0.727	2.572	0.727	2.550	0.728	2.558	0.728	2.564	0.727	2.565	0.728	2.565	0.727	2.562	0.728
	Slope	-0.011	0.023	-0.011	0.023	-0.011	0.023	-0.011	0.023	-0.010	0.023	-0.011	0.023	-0.011	0.023	-0.011	0.023	-0.011	0.023	-0.011	0.023
Resilience	Intercept	0.921	1.050	0.981	1.049	0.935	1.052	0.913	1.050	0.862	1.052	0.956	1.049	0.901	1.050	0.929	1.051	0.95	1.048	1.007	1.049
	Slope	0.441	0.354	0.421	0.354	0.437	0.355	0.444	0.354	0.462	0.355	0.430	0.354	0.448	0.354	0.439	0.354	0.432	0.353	0.413	0.354
Hope Agency	Intercept	1.567	0.773	1.596	0.776	1.657	0.778	1.652	0.778	1.642	0.779	1.656	0.779	1.630	0.773	1.606	0.776	1.634	0.778	1.663	0.777
	Slope	0.052	0.061	0.050	0.061	0.045	0.061	0.045	0.061	0.046	0.061	0.045	0.061	0.047	0.060	0.049	0.061	0.047	0.061	0.045	0.061
Hope Pathway	Intercept	1.963	0.897	1.957	0.896	1.970	0.897	1.976	0.898	1.952	0.895	1.965	0.895	1.963	0.897	1.971	0.895	1.981	0.899	1.977	0.895
	Slope	0.022	0.073	0.022	0.073	0.021	0.073	0.021	0.073	0.023	0.073	0.022	0.073	0.022	0.073	0.021	0.073	0.021	0.073	0.021	0.073

Table 5.8: Univariate Logistic Regression of Progression by Personal Attribute Scales at Baseline

Personal Attribute	Complete Case Slope Estimate	Combined Slope Estimate $\bar{\theta}_D$	Within-imputation Variance \bar{W}_D	Between-imputation Variance \bar{B}_D	Total Variance \bar{T}_D	95% Confidence Interval
Mindset	-0.001	0.001	0.013	1.57e-5	0.013	(-0.224, 0.227)
Self Efficacy	-0.001	0.000	0.001	2.71e-6	0.001	(-0.063, 0.062)
Self Esteem	-0.005	-0.011	0.001	1.00e-7	0.001	(-0.056, 0.034)
Resilience	0.661	0.437	0.125	1.91e-4	0.126	(-0.258, 1.131)
Hope Agency	0.044	0.047	0.004	6.10e-6	0.004	(-0.072, 0.167)
Hope Pathway	0.018	0.022	0.005	4.89e-7	0.005	(-0.121, 0.165)

Table 5.9: Combined Estimates and Sampling Variability by Personal Attribute Scales at Baseline

Using Generalized Likelihood Ratio Tests with forward selection and backwards elimination the model in Sex alone was described as the best model for continuation by all the imputed dataset. When looking at AIC, all of the imputed datasets suggest that the model in Sex + Age + Faculty best describes continuation, while BIC for all of the imputed datasets suggested that the null model best describes continuation. As in Section 5.2.1 the model in Sex and the model in Sex + Age + Faculty are the same for all of the imputed datasets. These preferred models contain none of the personal attribute scales.

Table 5.10 shows the model for Sex alone and the additive model for Sex, Age and Faculty. A model in Sex and Age and Faculty, that included interaction terms among the variables, was fitted. However, the interactions were not significant, with all p-values for interaction terms greater than our significance level of 0.05.

Models	Coef	Std Error	P-value	df	Residual Deviance	AIC	BIC
Sex							
Intercept	2.472	0.152	<0.001				
Sex (Male)	-0.561	0.218	0.010	967	610.61	614.61	624.37
Sex + Age + Faculty							
Intercept	1.940	0.274	<0.001				
Sex (Male)	-0.510	0.220	0.020				
Age (Under 21)	0.528	0.273	0.053				
Faculty (Profession)	0.478	0.335	0.153	965	604.68	612.68	632.19

Table 5.10: Logistic Regression of Progression for Sex and for Sex & Age & Faculty at Baseline

For the model in Sex alone the odds ratio for Sex is 0.57 with a confidence interval of (0.37, 0.87), signifying that the odds on a female student progressing with their original degree programme at the University of Glasgow after 1st year are between 1.15 and 2.70 times higher than a male student.

The p-value for Age is 0.053, which is slightly greater than our significance level of 0.05 indicating that Age is not statistically significant related to whether a student progresses when Sex and Faculty is included in the model. Faculty is not a significant predictor of progression as Table 5.10 shows the p-value for Faculty is greater than our significance level of 0.05.

Sex has an odds ratio of 0.60 with a confidence interval of (0.39, 0.92), indicating that female students have between 1.09 and 2.56 time higher odds of progression at the University of Glasgow than male students for any given faculty and age group.

When comparing the model building results above to the complete case analysis in Section 4.2 and Table 4.8 it can be seen that, when using BIC, both the complete case and the full dataset analysis chose the null model. In the complete case analysis the additive model in Sex and Age was chosen as the best model for progression by Generalized Likelihood Ratio Tests and AIC. For the full dataset Generalized Likelihood Ratio Tests and AIC have gone different ways with AIC adding Faculty and GLRT dropping Age. From Table 4.7 it can be seen that the significant results were marginal between Sex + Age and Sex + Age + Faculty for AIC and between Sex + Age and Sex for the Generalized Likelihood Ratio Test. Unlike continuation there is less agreement between the 3 methods on which model is best when more data is available for the model building process.

5.3 Imputing at Baseline, Semester 1 and Semester 2

The purpose of this multiple imputation is to obtain a full set of data for every student. As the study only asked students to fill out a follow-up questionnaire at Semester 1 or Semester 2, even if every student had filled in every single item at Baseline and at Semester 1 or Semester 2 the dataset would still have half of the Semester 1 and half of the Semester 2 values missing. This would too much for the **mi package** to compute. Therefore the Semester 1 and Semester 2 results were combined with a semester indicator variable, similar to Section 4.3 and Section 4.4.

It was decided that imputation should be conducted so that each student had personal attributes scores for 1 follow up visit and for Baseline. To obtain this the imputations were split into two steps where first step imputed missing items and the second step imputed missing scale items. This was split into two steps as the computing requirements for doing both steps at once were too enormous.

Step 1

The first step was to impute the missing items at Baseline and at Semester 1/Semester 2 for students that attempted a follow up questionnaire. The Semester 1 and Semester 2 items were combined and a binary semester variable was created to establish the semester in which the student attempted a follow up questionnaire. The items were imputed separately for each personal attribute scale. For Hope, the 12 items were imputed together instead of splitting into a Hope Agency imputation and a Hope Pathway imputation. The predictors included for each imputation model were Sex, Age, Faculty, Domicile, SEC, the baseline personal attribute items, the follow up personal attribute items and the semester indicator. For each of the imputations the imputed item values were treated as ordered categorical variables.

For each personal attribute 10 imputations were conducted with a maximum of 1000 iterations to establish if the imputation had converged. Imputed dataset were constructed so that every student that attempted a follow up questionnaire had a value for every item at Baseline and values for every item at Semester 1/Semester 2. This was done by randomly selecting, without replacement, one of the imputed datasets for each personal attribute and combining them into a full dataset. This resulted in 10 imputed datasets and the personal attribute scale scores were calculated in each.

Step 2

Once the follow up items had been imputed for students that attempted a follow up questionnaire and the personal attribute scales' scores had been calculated, the next step was to impute a scale value for students that did not attempted a follow up questionnaire. The dataset used by the mi package was created by randomly selecting one of the datasets (dataset 3) from step 1 and combining with the imputed item values for students that did not attempt a follow up questionnaire from a dataset (dataset 4) imputed in section 5.2. This was done so that students that did attempt a follow up questionnaire had a complete set of data and students that did not attempt a follow up questionnaire had a complete set of baseline data.

It was decided that the scale values would be imputed for each personal attribute separately as in Step 1. For each of the imputation models the predictors included are Sex, Age, Faculty, Domicile, SEC, the baseline scale score, the follow up scale score and the semester indicator.

The semester in which the student was invited to complete the follow up questionnaire was not produced for 10 students. This was because 2 of the students had not given consent for them to be contacted again while 8 of the students had withdrawn from the university before the semester allocation had been generated. It was decided to still include these students in the dataset to be imputed and randomly generate a semester for them.

10 imputations were conducted with each having a maximum of 1000 iterations to establish if the imputation has converged. Once the 10 imputed datasets for each personal attribute scale had been produced they were used to construct 10 complete datasets each with a baseline and follow up scales score for every personal attribute scale. This was done by using the same method in step 1 of combining 1 randomly selected imputed dataset for each of the personal attribute scales. It is these datasets that will now be examined.

5.3.1 Continuation for Imputed Data at Baseline and Semester 1/ Semester 2 for Difference in Personal Attribute Scores

The relationship of Continuation and the difference in each personal attribute score at Baseline and Semester 1/Semester 2 for each of the imputed datasets shall now be examined. The paired differences in Semester 1 and Semester 2 have been combined for each dataset as described in Section 4.3.

The individual univariate personal attribute models were fitted with and without the binary variable indicating in which Semester the student was asked to fill out the questionnaire. However it was established that the semester indicator variable was not statistically significant and it will be removed for the rest of the analysis described here.

The univariate analysis for the 5 demographic variables is described in section 5.2.1. The univariate analysis of each individual personal attribute scale was examined for all 10 imputed datasets. Table 5.11 contains the results of the univariate logistic regression analysis in each individual personal attribute, where the coefficient and standard error have been

highlighted in bold if the p-value is less than our significance level of 0.05. From this it can be seen that the difference in Self Esteem in dataset 5 (p-value = 0.014) and the difference in Hope Agency in dataset 7 (p-value = 0.006) individually are significant predictors of continuation. Looking at the slope estimates, in particular difference in Resilience, there is little consistency across the imputations. This could be because of the high amount of values that were imputed.

Table 5.12 shows that for each of the personal attributes the combined slope estimate for the imputed datasets is not similar to the complete case analysis estimate. Looking at the sampling variability it can be seen that the between-imputation variance is still quite small for each of the personal attributes apart from Resilience. Resilience has a high within-imputation and between-imputation variance indicating that there is less precision in the slope parameter estimate and that the parameter estimates are not consistent across the imputed datasets. However the confidence interval contains 0 suggesting that it is not significant. This is similar to the univariate logistic regression analysis for continuation at Baseline.

		1		2		3		4		5		6		7		8		9		10	
Personal Attribute		Coef	Std Error	Coef	Std Error	Coef	Std Error	Coef	Std Error	Coef	Std Error	Coef	Std Error	Coef	Std Error	Coef	Std Error	Coef	Std Error	Coef	Std Error
δMindset	Intercept	2.495	0.132	2.462	0.122	2.442	0.125	2.526	0.131	2.505	0.130	2.500	0.130	2.519	0.133	2.454	0.123	2.530	0.132	2.457	0.126
	Slope	0.092	0.156	-0.009	0.137	-0.088	0.153	0.234	0.165	0.149	0.162	0.119	0.156	0.151	0.143	-0.049	0.152	0.221	0.157	-0.023	0.149
δSelf Efficacy	Intercept	2.486	0.124	2.492	0.125	2.524	0.128	2.491	0.124	2.467	0.122	2.524	0.127	2.486	0.123	2.463	0.120	2.451	0.124	2.481	0.124
	Slope	0.031	0.038	0.037	0.038	0.068	0.039	0.039	0.038	0.006	0.039	0.073	0.039	0.037	0.039	-0.001	0.038	-0.026	0.039	0.021	0.039
δSelf Esteem	Intercept	2.485	0.124	2.464	0.122	2.512	0.127	2.477	0.122	2.552	0.130	2.488	0.125	2.470	0.123	2.500	0.126	2.476	0.122	2.461	0.121
	Slope	0.028	0.034	0.000	0.035	0.054	0.035	0.019	0.034	0.091	0.037	0.029	0.035	0.008	0.034	0.040	0.036	0.017	0.032	-0.006	0.034
δResilience	Intercept	2.477	0.125	2.438	0.124	2.429	0.123	2.443	0.122	2.393	0.122	2.459	0.122	2.439	0.120	2.464	0.124	2.471	0.127	2.449	0.120
	Slope	0.176	0.460	-0.320	0.466	-0.455	0.470	-0.294	0.429	-0.919	0.470	-0.102	0.467	-0.563	0.468	0.000	0.457	0.077	0.478	-0.670	0.438
δHope Agency	Intercept	2.476	0.126	2.456	0.123	2.430	0.121	2.504	0.130	2.444	0.123	2.456	0.124	2.617	0.140	2.432	0.123	2.472	0.126	2.483	0.130
	Slope	0.022	0.071	-0.015	0.062	-0.111	0.081	0.068	0.077	-0.047	0.081	-0.019	0.080	0.235	0.086	-0.080	0.083	0.016	0.076	0.032	0.080
δHope Pathway	Intercept	2.487	0.124	2.463	0.122	2.435	0.120	2.455	0.121	2.480	0.123	2.496	0.125	2.475	0.122	2.452	0.120	2.462	0.122	2.495	0.124
	Slope	0.067	0.078	-0.003	0.078	-0.156	0.085	-0.037	0.082	0.050	0.077	0.078	0.078	0.037	0.077	-0.060	0.077	-0.006	0.081	0.089	0.077

Table 5.11: Univariate Logistic Regression of Continuation by Difference in Personal Attribute Scales at Baseline and Semester 1/2

Personal Attribute	Complete Case Slope Estimate	Combined Slope Estimate $\bar{\theta}_D$	Within-imputation Variance \bar{W}_D	Between-imputation Variance \bar{B}_D	Total Variance \bar{T}_D	95% Confidence Interval
δMindset	-0.049	0.080	0.023	0.013	0.038	(-0.344, 0.504)
δSelf Efficacy	0.105	0.028	0.001	0.001	0.002	(-0.080, 0.137)
δSelf Esteem	0.130	0.028	0.001	0.001	0.002	(-0.072, 0.128)
δResilience	-0.628	-0.307	0.212	0.123	0.348	(-1.592, 0.978)
δHope Agency	0.121	0.010	0.006	0.009	0.016	(-0.272, 0.292)
δHope Pathway	0.018	0.006	0.006	0.006	0.013	(-0.241, 0.253)

Table 5.12: Combined Estimates and Sampling Variability by Difference in Personal Attribute Scales at Baseline and Semester 1/2

As in section 5.2.1 the same model building analysis will be used on all of the imputed datasets.

Imputed Dataset	GLRT	AIC	BIC
1	Faculty	Sex + Faculty	Faculty
2	Faculty	Sex + Faculty	Faculty
3	Faculty + δ Self Efficacy + δ Hope Pathway	Sex + Faculty + δ Self Efficacy + δ Hope Pathway	Faculty
4	Faculty	Sex + Faculty	Faculty
5	Faculty + δ Self Esteem + δ Resilience	Sex + Faculty + δ Self Esteem + δ Resilience	Faculty
6	Faculty	Sex + Faculty + Age + δ Self Efficacy	Faculty
7	Faculty + δ Hope Agency	Sex + Faculty + δ Resilience + δ Hope Agency	δ Hope Agency
8	Faculty	Sex + Faculty	Faculty
9	Faculty	Sex + Faculty	Faculty
10	Faculty	Sex + Faculty + δ Resilience	Faculty

Table 5.13 : Results of Model Building for Continuation

Table 5.13 shows that when using the Generalized Likelihood Ratio Test with forward selection and backwards elimination, 7 out of the 10 imputed datasets suggested that the model in Faculty best describes continuation. Faculty is included in the three other suggested models.

It can also be seen that AIC chooses Sex + Faculty for half of the imputed datasets as the best model to describe continuation and the other half had Sex + Faculty as part of the model.

For 9 out of the 10 imputed datasets BIC chose Faculty as the best model to describe continuation.

From Table 5.13 it appears that occasionally some of the differences in personal attribute scores are suggested as being part of a model that describes Continuation. However, there is no consistent evidence that any of the differences in personal attribute scores are significantly related to Continuation across the datasets. Instead, as also seen in Table 5.11, the occasional dataset will have one of the differences in personal attribute score with a p-value of less than 0.05. The number of tests is quite large in the course of fitting models across all the imputed datasets, consequently some false positive results are expected.

Overall I would say that the three model building techniques are suggesting the models that were suggested for Continuation in Section 5.2.1.

The models in Faculty alone and in Sex + Faculty are shown in Table 5.7. Results for the model in the difference in Hope Agency are shown in Table 5.11 and 5.12, where the probability of continuation increases as the difference in Hope Agency score increases in the 7th imputed dataset.

5.3.2 Progression for Imputed Data at Baseline and Semester 1/ Semester 2 for Difference in Personal Attribute Scores

The same analysis used for continuation in section 5.3.1 will be repeated for progression. The individual univariate personal attribute models were fitted with and without the Semester variable. However, it was established that the semester indicator variable was not statistically significant and was removed for the rest of the progression analysis. Section 5.2.2 describes the univariate analysis for the 5 demographic variables.

From Table 5.14 it can be seen that the difference in Hope Pathway in dataset 3, the difference in Self Esteem in dataset 5 and the difference in Resilience in dataset 10 individually are significant predictors of continuation. As in continuation for the slope estimates, in particular the difference in Resilience, there is little consistency across the imputations.

For each of the personal attributes, Table 5.15 shows that the combined slope estimate for the imputed datasets is not similar to the complete case analysis estimate. The within-imputation variance and the between-imputation variance are still quite small for each of the personal

attribute apart from Resilience. This indicates that there is substantial consistency across the imputed datasets. For Resilience, the within-imputation and between-imputation variance is quite large indicating that there is less precision in the slope parameter estimate and that the parameter estimates are not consistent across the imputed datasets. However the confidence interval contains 0 suggesting that it is not significant.

		1		2		3		4		5		6		7		8		9		10	
Personal Attribute		Coef	Std Error	Coef	Std Error	Coef	Std Error	Coef	Std Error	Coef	Std Error	Coef	Std Error	Coef	Std Error	Coef	Std Error	Coef	Std Error	Coef	Std Error
δMindset	Intercept	2.221	0.116	2.212	0.109	2.198	0.112	2.250	0.115	2.224	0.114	2.249	0.117	2.274	0.120	2.207	0.110	2.243	0.115	2.235	0.116
	Slope	-0.031	0.140	-0.148	0.117	-0.139	0.139	0.081	0.148	-0.030	0.146	0.064	0.142	0.121	0.130	-0.133	0.138	0.048	0.141	0.014	0.135
δSelf Efficacy	Intercept	2.224	0.110	2.220	0.109	2.256	0.113	2.262	0.113	2.223	0.110	2.264	0.113	2.228	0.109	2.229	0.112	2.223	0.110	2.243	0.113
	Slope	-0.013	0.033	-0.021	0.033	0.034	0.035	0.044	0.034	-0.015	0.036	0.048	0.035	-0.007	0.035	-0.003	0.035	-0.016	0.035	0.016	0.035
δSelf Esteem	Intercept	2.262	0.113	2.232	0.111	2.270	0.114	2.247	0.111	2.302	0.117	2.259	0.113	2.260	0.114	2.260	0.114	2.225	0.109	2.257	0.112
	Slope	0.038	0.031	0.001	0.032	0.046	0.032	0.023	0.031	0.079	0.034	0.032	0.032	0.030	0.031	0.034	0.033	-0.011	0.027	0.035	0.028
δResilience	Intercept	2.254	0.114	2.221	0.114	2.193	0.112	2.190	0.110	2.173	0.112	2.212	0.109	2.206	0.109	2.216	0.111	2.229	0.115	2.216	0.109
	Slope	0.298	0.415	-0.109	0.421	-0.504	0.427	-0.708	0.379	-0.676	0.424	-0.422	0.422	-0.534	0.425	-0.223	0.413	-0.018	0.434	-0.914	0.397
δHope Agency	Intercept	2.285	0.117	2.254	0.113	2.209	0.111	2.247	0.116	2.235	0.114	2.206	0.111	2.301	0.120	2.197	0.111	2.226	0.113	2.266	0.119
	Slope	0.088	0.060	0.038	0.043	-0.061	0.073	0.030	0.069	0.008	0.073	-0.063	0.072	0.125	0.078	-0.086	0.076	-0.009	0.069	0.056	0.073
δHope Pathway	Intercept	2.230	0.110	2.220	0.110	2.201	0.109	2.220	0.109	2.234	0.111	2.226	0.111	2.220	0.109	2.214	0.109	2.223	0.110	2.248	0.112
	Slope	-0.005	0.071	-0.045	0.071	-0.174	0.078	-0.047	0.075	0.012	0.070	-0.016	0.071	-0.048	0.070	-0.098	0.070	-0.029	0.074	0.054	0.070

Table 5.14: Univariate Logistic Regression of Progression by Difference in Personal Attribute Scales at Baseline and Semester 1/2

Personal Attribute	Complete Case Slope Estimate	Combined Slope Estimate $\bar{\theta}_D$	Within-imputation Variance \bar{W}_D	Between-imputation Variance \bar{B}_D	Total Variance \bar{T}_D	95% Confidence Interval
δMindset	-0.243	-0.015	0.019	0.010	0.030	(-0.388, 0.357)
δSelf Efficacy	0.078	0.007	0.001	0.001	0.002	(-0.090, 0.104)
δSelf Esteem	0.169	0.031	0.001	0.001	0.002	(-0.057, 0.119)
δResilience	-0.511	-0.381	0.173	0.134	0.321	(-1.625, 0.863)
δHope Agency	0.107	0.013	0.005	0.005	0.010	(-0.208, 0.233)
δHope Pathway	-0.010	-0.040	0.005	0.004	0.009	(-0.253, 0.174)

Table 5.15: Combined Estimates and Sampling Variability by Difference in Personal Attribute Scales at Baseline and Semester 1/2

As in section 5.3.1 the same model building analysis will be used on all of the imputed datasets.

Imputed Dataset	GLRT	AIC	BIC
1	Sex	Sex + Age + Faculty	Null
2	Sex	Sex + Age + Faculty	Null
3	Sex + δ Self Esteem + δ Hope Pathway	Sex + Age + Faculty + δ Self Efficacy + δ Self Esteem + δ Hope Pathway	Null
4	Sex	Sex + Age + Faculty	Null
5	Sex + δ Self Esteem + δ Resilience	Sex + Age + Faculty + δ Self Esteem + δ Resilience	δ Self Esteem
6	Sex	Sex + Age + δ Self Efficacy	Null
7	Sex	Sex + Age + Faculty + δ Resilience + δ Hope Agency	Null
8	Sex	Sex + Age + Faculty + δ Mindset + δ Self Esteem	Null
9	Sex	Sex + Age + Faculty	Null
10	Sex	Sex + Age + Faculty + δ Resilience	Null

Table 5.16: Results of Model building for Progression

Table 5.16 shows that, when using the Generalized Likelihood Ratio Test with forward selection and backwards elimination, 8 out of the 10 imputed datasets suggested that the model in Faculty best describes progression. The other 2 suggested models have Sex included in the models.

It can also be seen that AIC chooses Sex + Age + Faculty for only 4 out of the 10 imputed datasets as the best model to describe progression. However Sex + Age + Faculty is the only model that was selected in more than one dataset. The differences in personal attribute scores that have been included in the models suggested by AIC are not appearing consistently across the datasets. Instead it looks like for each dataset, it is choosing a different personal attribute.

For 9 out of the 10 imputed datasets BIC chose the null model as the best model to describe progression.

As in Continuation, from Table 5.16 it appears as though occasionally some of the differences in personal attribute score are suggested as being part of a model that describes Progression. There is no consistent evidence for this across the imputed datasets. Instead, as also seen in Table 5.11, the occasional dataset will have one of the differences in personal attribute score with a p-value of less than 0.05. The number of tests conducted in the course of fitting models across all the imputed datasets is quite large, consequentially some false positive results are expected. The relationship difference in Self Esteem and Progression that was seen in Section 4.4 appears to have disappeared with the more data that is now available in each dataset. The exception is imputed dataset 5 where the p-value is less than our significance level of 0.05

Overall I would say that the three model building techniques are suggesting the models that were suggested for Progression in Section 5.2.2.

The models in Sex alone and in Sex + Age + Faculty are shown in Table 5.10. Results for the model in the difference in Self Esteem are shown in Table 5.14 and 5.15, where the probability of progression increases as the difference in Self Esteem score increases in the 5th imputed dataset.

From the logistic regression analyses performed in this chapter it was found that none of the personal attribute scores were related to whether first year students continued or progressed at the University of Glasgow after first year; instead only Sex, Age and Faculty were suggested repeatedly as significant predictors. For the imputed datasets where follow up scale scores were imputed, occasionally some of the differences in personal attribute scores would be suggested as significant predictors of Continuation or Progression. However, there was no consistent evidence for this across the imputed datasets. A large number of statistical tests were conducted in the course of fitting models across all the imputed datasets, consequently some false positive results were to be expected.

Chapter 6

Discussion

6.1 Conclusions

The University of Glasgow set up this study to explore the relationship between the outcome in first year and students' personal attributes on entry to university and the changes in these attributes during first year. This is part of a programme of action that the university is currently engaged in to reduce the proportion of 1st year students who withdraw from the university during their first year. As the main interest of this study is the proportion of 1st year students who withdraw from the university during their first year, from exploring the data it became clear that the exclusion criteria the research team had applied had not captured a true representation of the population of 1st year students and that further exclusion criteria, arising out of a clearer definition of the target population needed to be implemented. The undergraduate students who had already obtained a degree before entering another course of study were excluded. However, it was not possible to exclude all undergraduate students who may have previously started a degree but whose credit did not count towards entry into their current degree at the University of Glasgow. This could possibly lead to the intended study population not being represented.

The completeness of the questionnaire returns has been documented in detail. The item non-response was recorded first to ascertain if any particular item or scale had been missed out or purposely not been answered by students. There was an impression that there was no pattern

of missingness for item non-response as the percentages of non-response within each scale were fairly evenly spread across the items. Overall at baseline, the amount of item non-response was small and even the percentage of missing scale values was no more than 4.8% for undergraduates (though it was as high as 8.7% for postgraduates). The percentages of missing personal attribute scales increased at Semester 1 and Semester 2 from Baseline; in particular, Self Efficacy and Resilience appeared to have a higher percentage of missing values consistently across both semesters.

Formal hypothesis tests were used to examine whether or not any demographic variables appeared to be related to non-completion of the survey by those who attempted it. First the proportion of missing scale scores, for each personal attribute, was analysed for the 5 demographic variables: Sex, Age, Faculty, Domicile and SEC for Undergraduates and Postgraduates for each time point. The only statistically significant result was for Resilience and Sex where a higher percentage of males than females failed to complete the Resilience scale items.

From this it was decided to look at whether the proportion of incomplete questionnaires was related to the 5 demographic variables. This was done for Undergraduates and Postgraduates at Baseline, Semester 1 and Semester 2. For Undergraduates at Baseline both Age and Faculty, individually, were statistically significantly related to the completeness of the questionnaire. None of the demographic variables was significantly related to the completeness of the questionnaire for Postgraduates. The Fisher's Exact Test showed that there was no statistically significant difference in the proportion of questionnaires that were completed and not completed between the sub-populations defined by any demographic variable for either Undergraduates or Postgraduates at Semester 1 or Semester 2. The missingness in Semester 1 and Semester 2 was not investigated in greater detail since the tests had little power due to the small number of responses and the even smaller number of missing responses at those time points.

Binary logistic regression and model building was implemented to further investigate the effects of the demographic variables on the completeness of the questionnaires. The univariate logistic regression results for Undergraduates agreed with the results from the Fisher's exact tests where only Age and Faculty separately are significant predictors of whether or not students completed the questionnaire. When selecting the model that best described completion, Generalized Likelihood Ratio Tests and AIC chose Age + Faculty and

BIC chose Faculty alone. When the additive model of Age + Faculty was fitted, Age was found to be statistically significant with Faculty included in the model.

For postgraduates the univariate logistic regression results also agreed with the Fisher's Exact test that none of the demographic variables were significant predictors of whether or not students completed the questionnaire. Generalized Likelihood Ratio Tests and BIC chose the null model and AIC chose the model in Age alone as the model that best describes whether or not students completed the questionnaire. However it had already been established that Age was not a statistically significant predictor of completion.

One of the aims in this thesis was to investigate the relationship between students' personal attributes and whether or not they continue and progress at the University of Glasgow after first year. Before logistic regression was used to analyse the data, the issue of missing data within the personal attributes needed to be overcome. Chapter 4 has described the logistic regression for a Complete Case analysis and Chapter 5 has described the same logistic regression for datasets filled using two different approaches to multiple imputation.

When investigating the differences in personal attribute scores at follow up Semester 1 and Semester 2 was combined with a semester variable indicator to account for any possible differences between the two semesters. The semester indicator variable was included in the model for multiple imputation. However, from the univariate logistic regression models in Chapter 4 and Chapter 5, it was established that the semester indicator variable was not statistically significant and was removed for the model building.

From these logistic regression analyses it was found that none of the personal attribute scores were related to whether first year students continued or progressed at the University of Glasgow after first year; instead only Sex, Age and Faculty were suggested repeatedly as significant predictors. For the imputed datasets where follow up scale scores were imputed, occasionally some of the differences in personal attribute scores would be suggested as significant predictors of Continuation or Progression. However, there was no consistent evidence for this across the imputed datasets. A large number of statistical tests were conducted in the course of fitting models across all the imputed datasets, consequently some false positive results were to be expected.

From the univariate analysis, the demographic variables Sex and Faculty individually, were significant predictors of Continuation. Sex and Age, individually, were significant predictors of Progression with p-values less than our significance level of 0.05. These models agreed with already well known results that females are more likely to succeed in their University studies than males and that students in professional degree programmes are more likely to continue at University and complete a degree than students in non-professional (or general) programmes (Higher Education Policy Institute 2009).

All of the models suggested by the model building methods, throughout the different types of datasets, were simple models. When comparing the results of the Complete Case analysis dataset with the imputed datasets, there was more agreement among the 3 criteria used for model building (GLRT, AIC, BIC) on which model is best for Continuation, with GLRT and BIC both suggesting a model in Faculty alone. However, for Progression the opposite occurred where AIC and GLRT no longer matched; instead of both suggesting the model in Sex + Age, as for the complete case analysis, GLRT now dropped Age and AIC added Faculty to the preferred model.

Throughout this thesis, there was speculation as to which model best described Continuation and Progression where AIC and BIC never agreed on the same model in any attempt of model building. BIC penalises larger models at a rate of $\log(n)$. This was the most stringent criterion and frequently suggested the null model. AIC was more flexible and regularly suggested larger models that included variables that were not statistically significant by the usual test. If GLRT didn't agree with AIC or BIC then it tended to suggest models that were intermediate between those suggested by AIC and BIC. My individual preference would be for AIC over BIC as BIC tended to favour the null model and it seems hard to believe that none of 11 explanatory variables had no impact on Continuation or Progression.

The multiple imputations that were carried out in this thesis were done using the **mi package** in **R**. Once I understood how the **mi.info** function worked and how to update the matrix of imputation information used for the imputations, I found that it was flexible and easy to update. It was especially helpful when using the ordered categorical option for imputing item values as the default model specifications had classed them as continuous. The only downside to it was that there was not an option to specify the maximum and minimum value for the imputation or at least not an option that I could find. Although this was not an issue for imputing item values as sensible item values were given, there were a few non sensible

values with scale level imputations. Over all for scale level imputations the majority of the values imputed were sensible which was very good considering that no limitations were imposed on the imputed values.

6.2 Limitations of the Study & Further Work

As all new entrants to the University were invited to take part in the study, everyone that took part was self selected. Within this every student that attempted a follow up questionnaire was again self selected. There is scope for potential self selection bias within the sample. The sample may not be a true representation of the population as the decision to participate in the study may reflect some inherent bias in the characteristics (including personal attributes) of the participating students. It is also unknown if the basic demographics samples are a true representation of the population as this data was not available. The response rate at Baseline was relatively low, especially for postgraduate students, and response rates for the follow up questionnaires were especially low. To investigate the potential for bias, the demographic information of the students that did not take part in the study could be compared with the demographic information of the students that did take part. It would also be of interest to ascertain whether, at follow up, missingness is related to scores obtained on one or more of the psychometric scales at Baseline to aid investigating the missing data mechanisms at work. Multiple imputation is dependent on the assumption that data is MAR. However it is very difficult to distinguish among MCAR, MAR, and NMAR for a given dataset. There is currently no test available to check that the MAR assumption holds. It could be possible that this assumption is not true.

The data collected on the 2009 cohort of students has been analysed in this thesis, but the analysis could be extend to more cohorts. The same data have been collected for the 2010 cohort of students and it may be of interest to apply the same techniques. It is also possible for this study to be replicated in another university, possibly in one located in the Glasgow area to compare the two separate universities within the same region.

As the personal attributes scores did not appear to have a significant effect on the outcomes of first year for students, perhaps future applications of this study could use alternative

psychometric scales to measure the personal attributes. It may be of interest to look into different personal attributes than the ones that were measured in this study.

Bibliography

- Akaike, H. (1973). '*Information theory as an extension of the maximum likelihood principle*', in B. N. Petrov & F. Csaki, (Eds.) '*Second International Symposium on Information Theory*' Akademiai Kiado, Budapest.
- Blascovich, J. & J. Tomaka (1993), '*Measures of Self-Esteem*' in J.P. Robinson, P.R. Shaver & L.S. Wrightsman (eds.), '*Measures of Personality and Social Psychological Attitudes*' (3rd ed.). Ann Arbor: Institute for Social Research.
- Bean, J.P. & S.B. Eaton (2000) '*A psychological model of college student retention*'. In: J.M. Braxton (ed. 2002) '*Reworking the student departure puzzle*' Nashville, Vanderbilt University Press. pp.48-61.
- Block, J. & A. M. Kerman (1996), '*IQ and Ego-Resiliency: Conceptual and Empirical Connections and Separateness*' , Journal of Personality and Social Psychology, Vol.70 No. 2, 349-361.
- Burnham, K. P. & D. R. Anderson (2004) '*Understanding AIC and BIC in Model Selection*', Sociological Methods Research, Vol. 33 No.2, 261-304.
- DeBerard, M. S., G. I. Spielmans & D. C. Julka (2004) '*Predictors of academic achievement and retention among college freshman*', College Student Journal, Vol. 38, No. 1, 66-80
- Dempster, A. P., N. M. Laird & D. B. Rubin (1977), '*Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion)*', Journal of the Royal Statistical Society, Series B 39, 1-38.
- Diggle, P. J., P. Heagerty, K.-Y. Liang, & S. L. Zeger (2002) '*Analysis of Longitudinal Data*' (2nd ed.), Oxford University Press Inc, Oxford.
- Dweck, C. S. (2000), '*Self-theories: Their role in motivation, personality and development*', Psychology Press, Philadelphia.
- Dweck, C.S, C. Chui & Y. Hong (1995) '*Implicit Theories and their role in judgements and reactions: A world from two perspectives*', Psychological Inquiry, Vol. 6 No. 4, 267-285.
- Gelman, A. & J. Hill (2006) '*Data Analysis Using Regression and Multilevel/Hierarchical Models*', Cambridge University Press, Cambridge.
- Higher Education Policy Institute (2009) '*Male and Female Participation and Progression in Higher Education*', HEPI, Oxford.
- Hosmer, D.W. & S. Lemeshow (1989), '*Applied Logistic Regression*', John Wiley and Sons Inc, New York.
- Fisher, R.A. (1925) '*Statistical Methods for Research Workers*' Oliver and Boyd, Edinburgh.

Kenward, M. –G. & J. Carpenter (2007) '*Multiple Imputation: current perspectives.*' Statistical Methods in Medical Research, 16; 199-218.

Little, R. J. A. & D. B. Rubin (2002), '*Statistical Analysis with Missing Data*' (2nd ed.), John Wiley and Sons Inc, New York.

Rosenberg, M. (1965), '*Society and the Adolescent Self-Image*' , Princeton University Press, Princeton, NJ.

Rosenberg, M. (1989), '*Society and the Adolescent Self-Image*' (Revised ed.), Wesleyan University Press, Middletown.

Rubin, D.B. (1987), '*Multiple Imputation for non response in surveys*', John Wiley and Sons Inc ,New York.

Schafer, J. L. (1997), '*Analysis of Incomplete Multivariate Data*' Chapman and Hall, London.

Schwarz, G. (1978). '*Estimating the Dimension of a Model*' Annals of Statistics, Vol. 6 No. 2, 461-464.

Schwarzer, R. & M. Jerusalem (1995). '*Generalized Self-Efficacy scale*', In J. Weinman, S. Wright, & M. Johnston, '*Measures in health psychology: A user's portfolio. Causal and control beliefs*' (pp. 35-37), Windsor, UK: NFER-NELSON.

Snyder, C.R., C. Harris, J.R. Anderson, S.A. Holleran, L.M. Irving, S.T. Sigmon et al. (1991), '*The Will and the Ways: Development and Validation of an Individual-Differences Measure of Hope*', Journal of Personality and Social Psychology, Vol. 60 No. 4, 570-585.

Snyder, C.R., H.S. Shorey, J. Cheavens, K.M. Pulvers, V.H. Adams III & Wiklund, C, (2002) '*Hope and academic success in college*', Journal of Educational Psychology, Vol. 94 No. 4, 820-826.

Su, Y. –S., A. Gelman, J. Hill & M. Yaima (2009). '*Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box*', Journals of Statistical Software, Vol. 45, Issue 2.

Web Resources

Higher Education Statistics Agency (www.hesa.ac.uk).

Schwarzer, R. (2011) '*Everything you wanted to know about the General Self-Efficacy Scale but were afraid to ask*', <http://www.ralfschwarzer.de/>

Weisstien, E. W. '*Fisher's Exact Test.*' From MathWorld – A Wolfram Web Resource, <http://mathworld.wolfram.com/FishersExactTest.html>.

Appendix A

Questionnaire

A.1 Personal Attributes Questionnaire

Below is the questionnaire that students were invited to complete at Baseline and at Semester 1 or Semester 2. Section 1 contains the psychometric scales for Mindset, Section 2 contains the psychometric scales for Self Efficacy, Section 3 contains the psychometric scales for Self Esteem, Section 4 contains the psychometric scales for Resilience and Section 5 contains the psychometric scales for Hope.

Survey of New Entrants

Attitudes to Learning and Transition to HE

Thank you for your willingness to take part in this survey. The questions aim to assess students' attitudes to learning and transition to studying at university level. It should only take you about 5 to 10 minutes to complete the questionnaire below. Please begin by filling in your Name and University of Glasgow Registration Number; this is a 7-digit number, of the form 09xxxxxx for students who are enrolling for the first time in September 2009. We need this information so we can identify you to contact you with feedback and in further stages of the study. We will also use your Registration Number to link to other information held by the University, such as the degree you are studying, so we don't have to ask you for this information again in this questionnaire. Be assured your responses will remain strictly confidential and will only be used by the researchers, they will not be revealed to anyone else in the University or outside. Information provided will be anonymised and will be treated in accordance with the Data Protection Act 1998. Participation in this survey is voluntary, though the more people who take part, the greater the likely benefit to students of the University.

[Note: Questions marked * must be completed]

1. * I consent to take part in this study

☐ Yes ☐ No

If you do not wish to take part, please close this window and return to pre-registration in WebSurf.

2. * Registration Number

This is the seven-digit number you used to log in to WebSurf


3. * Forename

4. * Surname

Section 1

Please indicate whether you agree or disagree with the statements below by selecting the most appropriate response from the drop down menu.

5. I have certain in-built talents, like sport or music, and I can't do much to change what those talents are.

Choose ... 

6. There are subjects, like maths or languages that I'm naturally good at, but others that I'm naturally poor at and I don't think I could ever be good in.

Choose ... ▼

7. To be honest, I don't think I can change how intelligent I am.

Choose ... ▼

8. Although I can learn new things, I can't really change what my talents and abilities are.

Choose ... ▼

Section 2

Please indicate whether you agree or disagree with the statements below by selecting the most appropriate response from the drop down menu.

9. I can always manage to solve difficult problems if I try hard enough.

Choose ... ▼

10. If someone opposes me, I can find the means and ways to get what I want.

Choose ... ▼

11. It is easy for me to stick to my aims and accomplish my goals.

Choose ... ▼

12. I am confident that I could deal efficiently with unexpected events.

Choose ... ▼

13. Thanks to my resourcefulness, I know how to handle unforeseen situations.

Choose ... ▼

14. I can solve most problems if I invest the necessary effort.

Choose ... ▼

15. I can remain calm when facing difficulties because I can rely on my coping abilities.

Choose ... ▼

16. When I am confronted with a problem, I can usually find several solutions.

Choose ... ▼

17. If I am in trouble, I can usually think of a solution.

Choose ... ▼

18. I can usually handle whatever comes my way.

Choose ... ▼

Section 3

Please indicate whether you agree or disagree with the statements below by selecting the most appropriate response from the drop down menu.

19. On the whole, I am satisfied with myself.

20. At times I think I am no good at all.

21. I feel that I have a number of good qualities.

22. I am able to do things as well as most other people.

23. I feel I do not have much to be proud of.

24. I certainly feel useless at times.

25. I feel that I am a person of worth, at least equal to others.

26. I wish I could have more respect for myself.

27. All in all, I am inclined to feel that I am a failure.

28. I take a positive attitude towards myself.

Section 4

Please indicate whether you agree or disagree with the statements below by selecting the most appropriate response from the drop down menu.

29. I am generous with my friends.

30. I quickly get over and recover from being startled.

31. I enjoy dealing with new and unusual situations.

32. I usually succeed in making a favorable impression on people.

33. I enjoy trying new foods I have never tasted before.

34. I am regarded as a very energetic person.

35. I like to take different paths to familiar places.

36. I am more curious than most people.

37. Most of the people I meet are likeable.

38. I usually think carefully about something before acting.

39. I like to do new and different things.

40. My daily life is full of things that keep me interested.

41. I would be willing to describe myself as a pretty 'strong' personality.

42. I get over my anger at someone reasonably quickly.

Section 5

Please indicate whether you agree or disagree with the statements below by selecting the most appropriate response from the drop down menu.

43. I can think of many ways to get out of a jam.

44. I energetically pursue my goals.

45. I feel tired most of the time.

46. There are lots of ways around any problem.

47. I am easily downed in an argument.

48. I can think of many ways to get the things in life that are important to me.

49. I worry about my health.

50. Even when others get discouraged, I know I can find a way to solve the problem.

51. My past experiences have prepared me well for my future.

52. I've been pretty successful in life.

53. I usually find myself worrying about something.

54. I meet the goals that I set for myself.

Clear Responses and Start Again

Finished - Send Questionnaire Off

(Have you filled in all the questions marked * ?)

Thank you very much for your time and co-operation. You will shortly receive an email to your University email account with more information on the survey you have just completed. We may contact you again for a follow-up survey later in the year; while we very much hope that you will continue to take part in this study, we emphasise again that participation is voluntary. If you have any queries about this study, please contact the researchers by emailing Alison Browitt at a.browitt@admin.gla.ac.uk

Appendix B

Programming Code

B.1 Example Model Building Code

```
model{
  for(i in 1:10){                                #Repeat for each of the 10 Imputed Dataset

    # Create Empty dataset to store results
    tab<- matrix(0,11,5 ,
                  dimnames=list(NULL,c("Name", "Deviance", "AIC", "BIC", "Variables")))
    tab<-data.frame(tab)

    #####
    ###                                8 steps for each model                                ##
    #####

    #1. Fit model.                                #2. Assign Model Unique Name.
    #3. Assign Row Number in tab dataset. #4. Output Model variables.
    #5. Output Deviance.                    #6. Output AIC
    #7. Output BIC.                        #8. Assign No. of Variables in Model

    ###Start with Null Model
    mod<-glm(as.factor(continuation)~ 1,family=binomial,data=d.list[[i]])
    assign(paste("modnull"i,sep="."),mod)
    x<-1
    tab[x,1]<-"null"
    tab[x,2]<-mod.null$deviance
    tab[x,3]<-mod.null$aic
    tab[x,4]<-mod.null$deviance + (n.c - mod.null$df.residual)*lognc    ##BIC
    tab[x,5]<-0

    ###For One Variable
    for(k in 1:10){
      mod<-glm(formula(paste("continuation ~",varnames$var.name[k],sep="")),
                family=binomial,data=d.list[[i]])

      assign(paste("mod",varnames[k,3],i,sep="."),mod)
      x<-x+1
      tab[x,1]<-paste(varnames[k,3],sep=".")
      tab[x,2]<-mod$deviance
      tab[x,3]<-mod$aic
    }
  }
}
```

```

    tab[x,4]<-mod$deviance + (n.c - mod$df.residual)*lognc ##BIC
    tab[x,5]<-1
  }

  ###For Two Variable
  for(l in 1:9){
    for(k in 2:10){
      if(l<k){
        mod<-glm(formula(paste("continuation ~",varnames$var.name[l],"+",
varnames$var.name[k],sep="")),family=binomial,data=d.list[[i]])

        assign(paste("mod",varnames[l,3],varnames[k,3],i,sep="."),mod)
        x<-x+1
        tab[x,1]<-paste(varnames[l,3],varnames[k,3],sep=".")
        tab[x,2]<-mod$deviance
        tab[x,3]<-mod$aic
        tab[x,4]<-mod$deviance + (n.c - mod$df.residual)*lognc ##BIC
        tab[x,5]<-2
      }
    }
  }

  ###For Three Variable
  for(m in 1:8){
    for(l in 2:9){
      for(k in 3:10){
        if(m<l &l<k){
          mod<-glm(formula(paste("continuation ~",varnames$var.name[m],"+",
varnames$var.name[l],"+",varnames$var.name[k],sep="")),
family=binomial,data=d.list[[i]])

          assign(paste("mod",varnames[m,3],varnames[l,3],
varnames[k,3],i,sep="."),mod)
          x<-x+1
          tab[x,1]<-paste(varnames[m,3],varnames[l,3],varnames[k,3],sep=".")
          tab[x,2]<-mod$deviance
          tab[x,3]<-mod$aic
          tab[x,4]<-mod$deviance + (n.c - mod$df.residual)*lognc ##BIC
          tab[x,5]<-3
        }
      }
    }
  }

  ###For Four Variable
  for(n in 1:7){
    for(m in 2:8){
      for(l in 3:9){
        for(k in 4:10){
          if(n<m &m<l &l<k){
            mod<-glm(formula(paste("continuation ~",varnames$var.name[n],"+",
varnames$var.name[m],"+",varnames$var.name[l],"+",
varnames$var.name[k],sep="")),family=binomial,data=d.list[[i]])

            assign(paste("mod",varnames[n,3],varnames[m,3],
varnames[l,3],varnames[k,3],i,sep="."),mod)
            x<-x+1
            tab[x,1]<- paste(varnames[n,3],varnames[m,3],varnames[l,3],
varnames[k,3],sep=".")
            tab[x,2]<-mod$deviance
            tab[x,3]<-mod$aic
            tab[x,4]<-mod$deviance + (n.c - mod$df.residual)*lognc ##BIC

```

```

        tab[x,5]<-4
    }
}
}}}}

###For Five Variable
for(o in 1:6){
  for(n in 2:7){
    for(m in 3:8){
      for(l in 4:9){
        for(k in 5:10){
          if(o<n & n<m & m<l & l<k){
            mod<-glm(formula(paste("continuation ~",varnames$var.name[o],"+",
              varnames$var.name[n],"+",varnames$var.name[m],"+",
              varnames$var.name[l],"+",varnames$var.name[k],sep="")),
              family=binomial,data=d.list[[i]])

            assign(paste("mod",varnames[o,3],varnames[n,3],
              varnames[m,3],varnames[l,3],varnames[k,3],i,sep="."),mod)
            x<-x+1
            tab[x,1]<-paste(varnames[o,3],varnames[n,3],varnames[m,3],
              varnames[l,3],varnames[k,3],sep=".")
            tab[x,2]<-mod$deviance
            tab[x,3]<-mod$aic
            tab[x,4]<-mod$deviance + (n.c - mod$df.residual)*lognc ##BIC
            tab[x,5]<-5
          }
        }
      }
    }
  }
}
}}}}

###For Six Variable
for(p in 1:5){
  for(o in 2:6){
    for(n in 3:7){
      for(m in 4:8){
        for(l in 5:9){
          for(k in 6:10){
            if(p<o & o<n & n<m & m<l & l<k){
              mod<-glm(formula(paste("continuation ~",varnames$var.name[p],"+",
                varnames$var.name[o],"+",varnames$var.name[n],"+",
                varnames$var.name[m],"+",varnames$var.name[l],"+",
                varnames$var.name[k],sep="")),family=binomial,data=d.list[[i]])

              assign(paste("mod",varnames[p,3],varnames[o,3],
                varnames[n,3],varnames[m,3],varnames[l,3],
                varnames[k,3],i,sep="."),mod)
              x<-x+1
              tab[x,1]<-paste(varnames[p,3],varnames[o,3],varnames[n,3],
                varnames[m,3],varnames[l,3],varnames[k,3],sep=".")
              tab[x,2]<-mod$deviance
              tab[x,3]<-mod$aic
              tab[x,4]<-mod$deviance + (n.c - mod$df.residual)*lognc ##BIC
              tab[x,5]<-6
            }
          }
        }
      }
    }
  }
}
}}}}

###For Seven Variable
for(q in 1:4){
  for(p in 2:5){
    for(o in 3:6){
      for(n in 4:7){
        for(m in 5:8){
          for(l in 6:9){

```

```

for(k in 7:10){
  if(q<p & p<o & o<n & n<m & m<l & l<k){
    mod<-glm(formula(paste("continuation ~",varnames$var.name[q],"+",
      varnames$var.name[p],"+",varnames$var.name[o],"+",
      varnames$var.name[n],"+",varnames$var.name[m],"+",
      varnames$var.name[l],"+",varnames$var.name[k],sep="")),
      family=binomial,data=d.list[[i]])

    assign(paste("mod",varnames[q,3],varnames[p,3],varnames[o,3],
      varnames[n,3],varnames[m,3],varnames[l,3],varnames[k,3],
      i,sep="."),mod)

    x<-x+1
    tab[x,1]<-paste(varnames[q,3],varnames[p,3],varnames[o,3],
      varnames[n,3],varnames[m,3],varnames[l,3],varnames[k,3],sep=".")
    tab[x,2]<-mod$deviance
    tab[x,3]<-mod$aic
    tab[x,4]<-mod$deviance + (n.c - mod$df.residual)*lognc ##BIC
    tab[x,5]<-7
  }
}

###For Eight Variable
for(r in 1:3){
  for(q in 2:4){
    for(p in 3:5){
      for(o in 4:6){
        for(n in 5:7){
          for(m in 6:8){
            for(l in 7:9){
              for(k in 8:10){
                if(r<q & q<p & p<o & o<n & n<m & m<l & l<k){
                  mod<-glm(formula(paste("continuation ~",varnames$var.name[r],
                    "+",varnames$var.name[q],"+",varnames$var.name[p],"+",
                    varnames$var.name[o],"+",varnames$var.name[n],"+",
                    varnames$var.name[m],"+",varnames$var.name[l],"+",
                    varnames$var.name[k],sep="")),family=binomial,data=d.list[[i]])

                  assign(paste("mod",varnames[r,3],varnames[q,3],varnames[p,3],
                    varnames[o,3],varnames[n,3],varnames[m,3],varnames[l,3],
                    varnames[k,3],i,sep="."),mod)

                  x<-x+1
                  tab[x,1]<-paste(varnames[r,3],varnames[q,3],varnames[p,3],
                    varnames[o,3],varnames[n,3],varnames[m,3],varnames[l,3],
                    varnames[k,3],sep=".")
                  tab[x,2]<-mod$deviance
                  tab[x,3]<-mod$aic
                  tab[x,4]<-mod$deviance + (n.c - mod$df.residual)*lognc ##BIC
                  tab[x,5]<-8
                }
              }
            }
          }
        }
      }
    }
  }
}

###For Nine Variable
for(s in 1:2){
  for(r in 2:3){
    for(q in 3:4){
      for(p in 4:5){
        for(o in 5:6){
          for(n in 6:7){
            for(m in 7:8){
              for(l in 8:9){
                for(k in 9:10){

```

```

if(s<r &r<q &q<p &p<o &o<n &n<m &m<l &l<k){
  mod<-glm(formula(paste("continuation ~",varnames$var.name[s],
    "+",varnames$var.name[r],"+",varnames$var.name[q],"+",
    varnames$var.name[p],"+",varnames$var.name[o],"+",
    varnames$var.name[n],"+",varnames$var.name[m],"+",
    varnames$var.name[l],"+",varnames$var.name[k],sep="")),
    family=binomial,data=d.list[[i]])

  assign(paste("mod",varnames[s,3],varnames[r,3],varnames[q,3],
    varnames[p,3],varnames[o,3],varnames[n,3],varnames[m,3],
    varnames[l,3],varnames[k,3],i,sep="."),mod)
  x<-x+1
  tab[x,1]<-paste(varnames[s,3],varnames[r,3],varnames[q,3],
    varnames[p,3],varnames[o,3],varnames[n,3],varnames[m,3],
    varnames[l,3],varnames[k,3],sep=".")
  tab[x,2]<-mod$deviance
  tab[x,3]<-mod$aic
  tab[x,4]<-mod$deviance + (n.c - mod$df.residual)*lognc ##BIC
  tab[x,5]<-9
}
}}}}}}}}

###Full Model
mod<- glm(as.factor(continuation)~ as.factor(Sex)+as.factor(Age.b)+
  as.factor(Dom)+ as.factor(faculty.b)+ diff.mindset+
  diff.selfefficacy+ diff.selfesteem+ diff.resilience+
  diff.hope.agency+diff.hope.pathway, family=binomial,data=d.list[[i]])

assign(paste("modfull",i,sep="."),mod)
x<-x+1
tab[x,1]<-"full"
tab[x,2]<-mod.all$deviance
tab[x,3]<-mod.all $aic
tab[x,4]<-mod.all$deviance + (n.c - mod.all$df.residual)*lognc ##BIC
tab[x,5]<-10

tab[,2:4]<-round(tab[2:4],2)

##Create the file name for the New Deviance, AIC & BIC data set then Output
fsavename<-paste("f:/New Comb Continuation Table",i,".csv",sep="")

write.csv(tab, file=fsavename)
}
}

```

B.2 Code for Imputing Scale level Data at Baseline

```
library(mi)

#Set up dataset to only include: Faculty, Sex, Age, Domicile, SEC,
# and the 5 personal attribute scale scores.
m1<-data.frame(d.ub[,c(7:8,10:11,16,72:76)])

#Information Matrix for Imputations
inf1<-mi.info(m1)
inf1

#Run imputations
imput <-mi(m1,info=inf1,n.imp=10,n.iter=1000,
          max.minutes=500,add.noise=noise.control(post.run.iter=50))

#Save Imputed Data Sets as csv files
write.mi(imput,format=c("csv"),row.names=F)
```

B.3 Code for Imputing Item level Data at Baseline

```
library(mi)

#Set up dataset for each personal attribute to only include:
#Faculty, Sex, Age, Domicile, SEC and the personal attribute items.
mindset.m<-dub.mi[,c(1:5,8:11)]
selfefficacy.m<-dub.mi[,c(1:5,12:21)]
selfesteem.m<-dub.mi[,c(1:5,22:31)]
resilience.m<-dub.mi[,c(1:5,32:45)]
hope.m<-dub.mi[,c(1:5,46:57)]

#Information Matrix for Imputations

mindset.i<-mi.info(mindset.m)
selfefficacy.i<-mi.info(selfefficacy.m)
selfesteem.i<-mi.info(selfesteem.m)
resilience.i<-mi.info(resilience.m)
hope.i<-mi.info(hope.m)

#Update Mindset to items are ordered catagorical
mindset.i<-update(mindset.i, "type", list(mindset1="ordered-categorical",
mindset2="ordered-categorical",mindset3="ordered-
categorical",mindset4="ordered-categorical"))

#Run imputations for each Personal Attribute Scale
mindset.im<-mi(mindset.m,info=mindset.i,n.imp=10,
  n.iter=1000,max.minutes=500,add.noise=noise.control(post.run.iter=50))

selfefficacy.im<-mi(selfefficacy.m,info=selfefficacy.i,n.imp=10,
  n.iter=1000,max.minutes=500,add.noise=noise.control(post.run.iter=50))

selfesteem.im<-mi(selfesteem.m,info=selfesteem.i,n.imp=10,
  n.iter=1000,max.minutes=500,add.noise=noise.control(post.run.iter=50))

resilience.im<-mi(resilience.m,info=resilience.i,n.imp=10,
  n.iter=1000,max.minutes=500,add.noise=noise.control(post.run.iter=50))

hope.im<-mi(hope.m,info=hope.i,n.imp=10,
  n.iter=1000,max.minutes=500,add.noise=noise.control(post.run.iter=50))

#Save Imputed Data Sets as csv files

write.mi(mindset.im,format=c("csv"),row.names=F)
write.mi(selfefficacy.im,format=c("csv"),row.names=F)
write.mi(selfesteem.im,format=c("csv"),row.names=F)
write.mi(resilience.im,format=c("csv"),row.names=F)
write.mi(hope.im,format=c("csv"),row.names=F)
```

B.4 Code for Imputing at Baseline and Semester – Step 1

```
library(mi)

#Read in Data
d.ucs<-read.csv("f:/my documents/project/ug ex/d.ucs.csv",header=T,
               na.strings =
list("#N/A","NA"))

#####
## Steps for each personal attribute ##
## 1. Create dataset to only include: Faculty, Sex, Age, Domicile, SEC, ##
## Semester indicator, the Baseline items and the follow up items. ##
## 2. Set up Information Matrix for Imputations. ##
## 3. Update Information matrix, if needed, so items are categorical. ##
## 4. Run Imputations. ##
## 5. Save Imputed Data Sets as csv files. ##
#####
##### Mindset #####
mindset.m<-d.ucs[,c(7:8,10:11,16,22:25,87:90,159)]
mindset.i<-mi.info(mindset.m)
mindset.i

mindset.i<-update(mindset.i, "type", list(Bmindset1="ordered-categorical",
    Bmindset2="ordered-categorical",Bmindset3="ordered-categorical",
    Bmindset4="ordered-categorical",Smindset1="ordered-categorical",
    Smindset2="ordered-categorical",Smindset3="ordered-categorical",
    Smindset4="ordered-categorical"))
mindset.i

mindset.im<-mi(mindset.m,info=mindset.i,n.imp=10,
    n.iter=1000,max.minutes=500,add.noise=noise.control(post.run.iter=50))

write.mi(mindset.im,format=c("csv"),row.names=F)

##### Self Efficacy #####
selfefficacy.m<-d.ucs[,c(7:8,10:11,16,26:35,91:100,159)]
selfefficacy.i<-mi.info(selfefficacy.m)
selfefficacy.i

selfefficacy.im<-mi(selfefficacy.m,info=selfefficacy.i,n.imp=10,
    n.iter=1000,max.minutes=500,add.noise=noise.control(post.run.iter=50))

write.mi(selfefficacy.im,format=c("csv"),row.names=F)

##### Self Esteem #####
selfesteem.m<-d.ucs[,c(7:8,10:11,16,36:45,101:110,159)]
selfesteem.i<-mi.info(selfesteem.m)
selfesteem.i

selfesteem.im<-mi(selfesteem.m,info=selfesteem.i,n.imp=10,
    n.iter=1000,max.minutes=500,add.noise=noise.control(post.run.iter=50))

write.mi(selfesteem.im,format=c("csv"),row.names=F)

##### Resilience #####
resilience.m<-d.ucs[,c(7:8,10:11,16,46:59,111:124,159)]
resilience.i<-mi.info(resilience.m)
resilience.i
```

```

resilience.im<-mi(resilience.m,info=resilience.i,n.imp=10,
  n.iter=1000,max.minutes=500,add.noise=noise.control(post.run.iter=50))

write.mi(resilience.im,format=c("csv"),row.names=F)

##### Hope #####
hope.m<-d.ucsf[,c(7:8,10:11,16,60:71,125:136,159)]
hope.i<-mi.info(hope.m)
hope.i

hope.im<-mi(hope.m,info=hope.i,n.imp=10,
  n.iter=1000,max.minutes=500,add.noise=noise.control(post.run.iter=50))

write.mi(hope.im,format=c("csv"),row.names=F)

#####
#####Combine the scales for a complete imputed data set#####
#####

#set file directory
setwd("f:/my documents/project/ug ex/overall imputed data sets")

#Create a Random Sample order for which imputed datasets should be combined
#Sequence of 1 to 10
x<-1:10
n.m <-sample(x)
n.ef<-sample(x)
n.es<-sample(x)
n.r <-sample(x)
n.h <-sample(x)

for (i in 1:10){

  #Set file names for the random datasets to be read in for each PA
  fname.mind<-paste("ug.mindset_",n.m[i],".csv",sep="")
  fname.eff<-paste("ug.selfefficacy_",n.ef[i],".csv",sep="")
  fname.estm<-paste("ug.selfesteem_",n.es[i],".csv",sep="")
  fname.res<-paste("ug.resilience_",n.r[i],".csv",sep="")
  fname.hope<-paste("ug.hope_",n.h[i],".csv",sep="")

  #Read in the random datasets
  mind<-read.csv(fname.mind,header=T)
  eff<-read.csv(fname.eff,header=T)
  estm<-read.csv(fname.estm,header=T)
  res<-read.csv(fname.res,header=T)
  hope<-read.csv(fname.hope,header=T)

  #Create new file by replacing the item values in the original dataset
  imp<-d.ucsf

  imp[,22:25]<-mind[,6:9]      ###Mindset Baseline
  imp[,26:35]<-eff[,6:15]     ###Self Efficacy Baseline
  imp[,36:45]<-estm[,6:15]    ###Self Esteem Baseline
  imp[,46:59]<-res[,6:19]     ###Resilience Baseline
  imp[,60:71]<-hope[,6:17]    ###hope Baseline
  imp[,87:90]<-mind[,10:13]   ###Mindset Semester
  imp[,91:100]<-eff[,16:25]   ###Self Efficacy Semester
  imp[,101:110]<-estm[,16:25] ###Self Esteem Semester
  imp[,111:124]<-res[,20:33]  ###Resilience Semester
  imp[,125:136]<-hope[,18:29] ###Hope Semester

```

```

####Calculate Personal Attribute Scores for Baseline and Follow up
imp$mindset.baseline<-(imp$Bmindset1+imp$Bmindset2+
                        imp$Bmindset3+imp$Bmindset4)/4

imp$selfefficacy.baseline<-(imp$Bselfefficacy1+imp$Bselfefficacy2+
                             imp$Bselfefficacy3+imp$Bselfefficacy4+imp$Bselfefficacy5+
                             imp$Bselfefficacy6+imp$Bselfefficacy7+imp$Bselfefficacy8+
                             imp$Bselfefficacy9+imp$Bselfefficacy10)

imp$selfesteem.baseline<-(imp$BSE.1+imp$BSE.2+imp$BSE.3+imp$BSE.4+
                          imp$BSE.5+imp$BSE.6+imp$BSE.7+imp$BSE.8+imp$BSE.9+imp$BSE.10)

imp$resilience.baseline<-(imp$Bresilience1+imp$Bresilience2+
                          imp$Bresilience3+imp$Bresilience4+ imp$Bresilience5+imp$Bresilience6+
                          imp$Bresilience7+imp$Bresilience8+imp$Bresilience9+imp$Bresilience10+
                          imp$Bresilience11+imp$Bresilience12+imp$Bresilience13+
                          imp$Bresilience14)/14
imp$resilience.baseline<-round(imp$resilience.baseline,1)

imp$hope.agency.baseline <-(imp$Bhope2+imp$Bhope9+imp$Bhope10+imp$Bhope12)
imp$hope.pathway.baseline <-(imp$Bhope1+imp$Bhope4+imp$Bhope6+imp$Bhope8)
imp$hope.total.baseline <-(imp$hope.agency.baseline
                          +imp$hope.pathway.baseline)

imp$SEM.mindset<-(imp$Smindset1+imp$Smindset2+
                  imp$Smindset3+imp$Smindset4)/4

imp$SEM.selfefficacy<-(imp$Sselfefficacy1+imp$Sselfefficacy2+
                       imp$Sselfefficacy3+imp$Sselfefficacy4+imp$Sselfefficacy5+
                       imp$Sselfefficacy6+imp$Sselfefficacy7+imp$Sselfefficacy8+
                       imp$Sselfefficacy9+imp$Sselfefficacy10)

imp$SEM.selfesteem<-(imp$SSE.1+imp$SSE.2+imp$SSE.3+imp$SSE.4+
                    imp$SSE.5+imp$SSE.6+imp$SSE.7+imp$SSE.8+imp$SSE.9+imp$SSE.10)

imp$SEM.resilience<-(imp$Sresilience1+imp$Sresilience2+imp$Sresilience3
                     +imp$Sresilience4+ imp$Sresilience5+imp$Sresilience6+
                     imp$Sresilience7+imp$Sresilience8+imp$Sresilience9
                     +imp$Sresilience10+imp$Sresilience11+imp$Sresilience12+
                     imp$Sresilience13+imp$Sresilience14)/14
imp$SEM.resilience<-round(imp$SEM.resilience,1)

imp$SEM.hope.agency<-(imp$Shope2+imp$Shope9+imp$Shope10+imp$Shope12)
imp$SEM.hope.pathway<-(imp$Shope1+imp$Shope4+imp$Shope6+imp$Shope8)
imp$SEM.hope.total<-(imp$SEM.hope.agency +imp$SEM.hope.pathway)

##Calculate Difference in score
imp$diff.mindset<-imp$SEM.mindset - imp$mindset.baseline
imp$diff.selfefficacy<-imp$SEM.selfefficacy - imp$selfefficacy.baseline
imp$diff.selfesteem<-imp$SEM.selfesteem - imp$selfesteem.baseline
imp$diff.resilience<-imp$SEM.resilience - imp$resilience.baseline
imp$diff.hope.total<-imp$SEM.hope.total - imp$hope.total.baseline
imp$diff.hope.agency<-imp$SEM.hope.agency - imp$hope.agency.baseline
imp$diff.hope.pathway<-imp$SEM.hope.pathway - imp$hope.pathway.baseline

##Create the file name for the new complete imputed data set then save
fsavename<-paste("f:/my documents/project/ug ex/overall imputed data sets
/ug.imp_",i,".csv",sep="")
write.csv(imp, file=fsavename)
}

```

B.5 Code for Imputing at Baseline and Semester – Step 2

```
#Read in Ug.com.comb.sem, This is the new combined data set
d.uscom<-read.csv("e:/My Documents/Project/UG Ex/Overall Imputed Data
Sets/ug.com.comb.sem.csv", header=T, na.strings = list("#N/A","NA"))

library(mi)

#####
## Steps for each personal attribute ##
## 1. Create dataset to only include: Faculty, Sex, Age, Domicile, SEC, ##
## Semester indicator, the Baseline Scale and the follow up scale. ##
## 2. Set up Information Matrix for Imputations. ##
## 3. Update Information matrix to identify student id number. ##
## 4. Run Imputations. ##
## 5. Save Imputed Data Sets as csv files. ##
#####

##### Mindset #####
mindset.m<-d.uscom[,c(1,7:8,10:11,16,72,137,160)]
mindset.i<-mi.info(mindset.m)
mindset.i
mindset.i<-mi.info.update.is.ID(mindset.i,list(IDrnum="TRUE"))
mindset.i

mindset.im<-mi(mindset.m,info=mindset.i,n.imp=10,
n.iter=100,max.minutes=500,add.noise=noise.control(post.run.iter=100))

write.mi(mindset.im,format=c("csv"),row.names=F)

##### Self Efficacy #####
selfefficacy.m<-d.uscom[,c(1,7:8,10:11,16,73,138,160)]
selfefficacy.i<-mi.info(selfefficacy.m)
selfefficacy.i
selfefficacy.i<-mi.info.update.is.ID(selfefficacy.i,list(IDrnum="TRUE"))
selfefficacy.i

selfefficacy.im<-mi(selfefficacy.m,info=selfefficacy.i,n.imp=10,
n.iter=100,max.minutes=500,add.noise=noise.control(post.run.iter=100))

write.mi(selfefficacy.im,format=c("csv"),row.names=F)

##### Self Esteem #####
selfesteem.m<-d.uscom[,c(1,7:8,10:11,16,74,139,160)]
selfesteem.i<-mi.info(selfesteem.m)
selfesteem.i
selfesteem.i<-mi.info.update.is.ID(selfesteem.i,list(IDrnum="TRUE"))
selfesteem.i

selfesteem.im<-mi(selfesteem.m,info=selfesteem.i,n.imp=10,
n.iter=100,max.minutes=500,add.noise=noise.control(post.run.iter=100))

write.mi(selfesteem.im,format=c("csv"),row.names=F)

##### Resilience #####
resilience.m<-d.uscom[,c(1,7:8,10:11,16,75,140,160)]
resilience.i<-mi.info(resilience.m)
resilience.i
resilience.i<-mi.info.update.is.ID(resilience.i,list(IDrnum="TRUE"))
resilience.i
```

```

resilience.im<-mi(resilience.m,info=resilience.i,n.imp=10,
  n.iter=100,max.minutes=500,add.noise=noise.control(post.run.iter=100))

write.mi(resilience.im,format=c("csv"),row.names=F)

##### Hope Agency #####
hopeage.m<-d.uscom[,c(1,7:8,10:11,16,77,142,160)]
hopeage.i<-mi.info(hopeage.m)
hopeage.i
hopeage.i<-mi.info.update.is.ID(hopeage.i,list(IDrnum="TRUE")) reference
hopeage.i

hopeage.im<-mi(hopeage.m,info=hopeage.i,n.imp=10,
  n.iter=100,max.minutes=500,add.noise=noise.control(post.run.iter=100))

write.mi(hopeage.im,format=c("csv"),row.names=F)

##### Hope Pathway #####
hopepath.m<-d.uscom[,c(1,7:8,10:11,16,78,143,160)]
hopepath.i<-mi.info(hopepath.m)
hopepath.i
hopepath.i<-mi.info.update.is.ID(hopepath.i,list(IDrnum="TRUE"))
hopepath.i

hopepath.im<-mi(hopepath.m,info=hopepath.i,n.imp=10,
  n.iter=100,max.minutes=500,add.noise=noise.control(post.run.iter=100))

write.mi(hopepath.im,format=c("csv"),row.names=F)

```