Using Predictive Risk Modelling to Identify Students at High Risk of Paper Non-Completion and Programme Non-Retention at University

Pengfei Jia

Department of Economics, Faculty of Business and Law

A dissertation submitted for the degree of Master of Business at Auckland University of Technology, Auckland, New Zealand

2014

Abstract

Course non-completion is of substantial concern to university and public funding bodies as it could potentially affect attrition rates and eventual educational performance. This paper seeks to empirically estimate the factors that affect paper non-completion and programme non-retention. More importantly, identifying students who are at high risk of course non-completion would provide opportunities for possible early intervention services. This study develops a predictive risk model (PRM) to estimate the likelihood of course non-completion among first-year students at a large public university in New Zealand. The main aim of this research is to explore the potential use of administrative data for targeting prevention and early interventions to university students. Our results suggest that many factors, including part-time study, ethnicity, gender, educational background, and programme study areas, could play a prominent role in predicting a student's risk of paper non-completion in the first year and non-retention in the second year at university. We assess the "target effectiveness" of our model from a number of perspectives. For example, the area under the ROC curves for paper non-completion and programme non-retention are 0.7553 and 0.7125, respectively. Students with the highest 10% of risk scores by our PRM would account for 29.25% of paper non-completions and 23.33% of programme non-retentions.

JEL Classification: I21

Keywords: Predictive risk modelling; Paper non-completion; Programme non-retention; University dropout

Contents

Tables and Figures

Acknowledgements

I would like to thank my supervisor Professor Tim Maloney, whose critical thinking, empirical expertise and dedication guided me through the whole process and helped me improve. I would also like to express my gratitude to the Department of Strategy and Planning at Auckland University of Technology for providing me with the administrative data used in this research. Any errors and omissions are my sole responsibility.

1. Introduction

University course non-completion[1] is of substantial concern in many countries over the past years due to its increasing frequency (Grubb, 1989; Hartog et al., 1989; Tinto, 1993; Montmarquette et al., 2001). This may be an increasing issue because of rising participation rates in university education over recent decades. As more young people study at university, failure rates might increase as less able or academically prepared students are admitted. In addition, public funding authorities are increasingly concerned by the potential waste of public expenditures on students who subsequently fail at university. As it could affect attrition rates and eventual educational performance, the understanding of a student's paper non-completion and programme non-retention behaviour is crucial to policy makers. For New Zealand higher education, the reduction of course non-completion is also at a core concern of recent reforms of tertiary education and it has been considered as an important factor for education quality assessment and funding (New Zealand Ministry of Education, 2004).

There is a substantial body of literature that has empirically examined the determinants of course non-completion (e.g., Wetzel et al., 1999; Montmarquette et al., 2001; Singell, 2004; Kerkvliet and Nowell, 2005; Bai and Maloney, 2006; Ishitani, 2006; Stratton et al., 2008; Belloc et al., 2010). Although a comprehensive understanding of the relative importance of various possible explanations for course non-completion remains elusive, it has been widely recognized that individual characteristics, student's educational background, and institutional characteristics are the main determinants of discontinuity or dropout behaviour (Robst et al.,

[1] In this research, we define both paper non-completion and programme non-retention as "course non-completion".

1998; Kerkvliet and Nowell, 2005). Nevertheless, mainly due to limited data availability, most previous studies have only utilized very few factors in their research. Using a more comprehensive dataset, our study is able to analyse the impact of a wide variety of explanatory variables on student's course non-completion.

Our paper makes contributions along this line by using a unique administrative dataset developed by the Department of Strategy and Planning at Auckland University of Technology (AUT). Our dataset includes detailed information about individual characteristics (e.g., ethnicity, country of origin, gender, first language status, part-time or full-time study), educational background (e.g., NCEA score, school decile, entrance type), and institutional or programme characteristics (e.g., average class size, paper size, level of complexity of the course, study area).

Administrative data are used in this research mainly because they have the advantage of being more complete and accurate than survey data, and are readily accessible to universities that might consider implementing this predictive tool for targeting services to at risk students. Also, administrative data are less likely to suffer from sample selection bias because we have access to the records of all first-year students enrolling at AUT University over our sample period. However, our administrative data do not include some potentially important background factors, such as financial and family factors which have been widely discussed in the literature (see Stampen and Cabrera, 1988; McPherson and Schapiro, 1991; Card and Lemieux, 2000; DesJardins et al., 2002; Kerkvliet and Nowell, 2005; Ishitani, 2006; Montmarquette et al., 2007; Stratton et al., 2008). Only a few studies have used administrative data to study student's non-completion or dropout behaviour (e.g., Robst et al., 1998; Singell, 2004; Bai and Maloney, 2006; Belloc et al. 2010).

In addition, as argued by Stratton et al. (2008), completion of first year appears to have the most significant impact on the likelihood of course non-completion. Our sample only includes first-year students who had enrolled in Bachelor degree programmes at AUT for the first time. Also, our study empirically analyses the student's paper non-completion and programme non-retention behaviour separately to gain additional intuition into poor academic outcomes at university.

Most importantly, none of the existing research has used available information to predict the overall inherent risks in academic outcomes for university students. Partial effects analysis can only enable us to explore the relative importance of specific risk factors for course non-completion. Policy makers at university, however, may be more interested in identifying students who are the most exposed to the risk of course non-completion at the outset of their initial enrolment as it may allow sufficient time for the early delivery of targeted prevention services (e.g., tutorials, workshops, student advising and mentoring services).

Utilizing predictive risk modelling (PRM), we are able to generate a risk score for every individual student. PRM has been successfully utilized in such areas as health care and child protection (e.g., see Billings et al., 2006, 2012; Vaithianathan et al., 2013). The use of predictive risk modelling has not, to our knowledge, been utilized previously in investigating the course completion behaviour of university students. Our empirical work contributes to the literature by providing the first such example. We develop our PRM based on maximum likelihood Probit regression analysis. The main objectives of this study are to gain general understanding of student paper non-completion and programme non-retention behaviour, and more importantly, and to explore the potential use of administrative data for targeting prevention and early intervention services to university students.

The rest of the paper is organized as follows. Section 2 provides an overview of the existing literature. We describe the data used to estimate the model in section 3. Section 4 presents the methods for estimation and robustness testing. Section 5 reports the estimation results, and section 6 concludes.

2. Literature review

There exists a substantial literature that seeks to explore the factors which account for student's discontinuity or dropout behaviour. Some of the key empirical studies are summarized in Table 1.

Table 1: Empirical Studies on Non-completion or Dropout Behaviour

Studies	Topic	Data	Estimation model	Main findings
Wetzel et al. (1999)	Factors affecting student retention probabilities	All freshman and sophomore students over 1989-1992 at an urban university in the U.S.	Logit model	1. Academic and social integration factors are the most important factors on retention. 2. The effects of financial factors are substantial, but less than academic factors. 3. There are distinct differences across white and black students on retention.
Montmarquette et al. (2001)	The determinants of university dropouts	A longitudinal dataset on student enrolments in 1987 and 1988 at	Bivariate Probit model	1. Academic performance is the key element in the dropout decision.

		the Université de Montréal in Canada		2. There are significant differences of demographic characteristics on student's dropout behaviour.
Singell (2004)	The effects of financial aid on retention	Administrative data for freshman applicants in 1997 and 1998 at the University of Oregon in the U.S.	Univariate and bivariate Probit model	1. Decision to drop out depends significantly on financial aid. 2. Other factors also affect retention, including academic performance and demographic characteristics.
Kerkvliet and Nowell (2005)	University differences in the influence of wages, financial aid, and integration on student retention	1,200 students surveyed at Weber State University and Oregon State University in 1997 in the U.S.	Negative binomial and Poisson model	1. Opportunity costs influence retention. 2. The effects of financial aid differ by university and are sometimes negative.
Bai and Maloney (2006)	Ethnicity and academic success at university	A unique dataset from a cohort of students who first entered a university in New Zealand in the 2000 academic year	OLS and Probit model	1. Ethic identification is a significant factor in explaining variations in university grades. 2. GPA is the single most important determinant of dropout behaviour.
Ishitani (2006)	Studying attrition and degree completion behaviour among first-generation college students	4,427 students who enrolled in institutions between 1991 and 1994 in the U.S. (from National Education Longitudinal Study)	Multiple logistic regression model	1. Race, gender, family income, high school rank, and financial aid have a significant influence on student's attrition.
Stratton et al. (2008)	A multinomial logit model of college stopout and dropout behaviour	A 5 year longitudinal dataset in which students were surveyed in 1989 and then were re-interviewed in 1994 in the U.S.	Multinomial Logit model	1. Personal, household, institutional, and economic factors could substantially affect stopout (i.e., temporary dropout) and dropout behaviour, but the influences differ between the two. 2. The effects of financial aid are significant, but the type of aid has a differential impact on stopout and dropout probabilities.
Belloc et al. (2010)	University dropout: An Italian experience	Administrative data on 9,725 undergraduates enrolled in	Generalized Linear Mixed model	1. Results show significant effects of individual's characteristics on dropout.

	Sapienza University of Rome in Italy from 2001 to 2007	2. Findings relate a high dropout probability to a high secondary school final mark and a low university academic performance.

Many studies have shown substantial differences in student demographic characteristics (ethnicity, gender, country of origin, age, etc.) on dropout or attrition behaviour (see Grayson, 1998; Robst et al., 1998; Wetzel et al., 1999; Montmarquette et al., 2001; Bai and Maloney, 2006; Mastekaasa and Smeby, 2008; Belloc et al., 2010; Rodgers, 2013). For example, using administrative data from a university in New Zealand, Bai and Maloney (2006) found that the average probabilities of dropout are 7.4 and 9.1 percentage points higher for Māori and Pacific Island students, compared to otherwise observationally equivalent students of other ethnicities. Similar evidence on ethnicity differences in retention is also provided by Wetzel et al. (1999) who have analysed the factors affecting the probabilities of retention using data from freshmen and sophomore students over the period 1989-1992. Findings show significant differences among white and black students in the impact of different factors on retention. In addition, some research indicates that the decision to drop out depends significantly on gender (Montmarquette et al., 2001, 2007; Mastekaasa and Smeby, 2008; Belloc et al., 2010). However, the results in this literature are highly inconsistent. Some studies have found that female students have a lower risk of paper non-completion compare to male students. (e.g., Montmarquette et al., 2001, 2007; Mastekaasa and Smeby, 2008), while Belloc et al. (2010) argued that female students are more likely to drop out. Including all the above demographic factors in our dataset, we also have additional information on personal characteristics, such as part-time or full-time study, first language status, domestic or international student standing, and the nature of the papers and programmes in which these students are enrolled.

It has been widely recognized that student's success at university is substantially affected by his or her prior academic performance (see Betts and Morell, 1999; Cohn et al., 2004; Cyrenne and Chan, 2012; Ficano, 2012). However, only few studies in the existing literature have empirically examined the impact of past academic performance on student's dropout behaviour (Wetzel et al., 1999; Montmarquette et al., 2001; Singell, 2004; Bai and Maloney, 2006; Ishitani, 2006; Stratton et al., 2008; Belloc et al., 2010; Ost, 2010). For students new to the university, Singell (2004) found that the marginal effects of a student's GPA on re-enrolling are positive, indicating that those with poor academic performance are more likely to drop out. Also, as in Montmarquette et al. (2001) and Bai and Maloney (2006), student's academic performance at university has been found to be a key determinant of dropout behaviour. Consistent with these results, Stratton et al. (2008) found that low GPA reduced the probability of continuous enrolment from 90.2% to 71.8% for students.

Having information about a student's pre-university academic records (i.e., NCEA score[2]), this study seeks to examine the role of high school academic performance in a student's first year course non-completion outcomes at AUT. We can also shed light on the relationship between high school socioeconomic ranking (using school deciles[3] as proxies) and non-completion behaviour since it has been found in prior work that students who come from low school deciles are more likely to drop out (see Bai and Maloney, 2006; Ishitani, 2006).

[2] The National Certificate of Educational Achievement (NCEA) is the official secondary school qualification in New Zealand. The NCEA system is made up of three certificates at Level-1, 2, and 3, which are unusually studied in years 11, 12, and 13. NCEA level 3 is the highest qualification that New Zealand students could have obtained by their last year of secondary school.

[3] A school decile indicates the socioeconomic status of the community from which a school draws its students. Decile 1 schools are the 10% of schools from the lowest socioeconomic areas. Decile 10 schools are the 10% of schools from the highest socioeconomic areas. The lower the decile ranking of a school, the more funding it receives to offset the learning disadvantages associated with deprivation.

Apart from the above factors, institutional characteristics, such as class size and the difficulty of the enrolled programme, could also play a role in a student's academic success (Tinto, 1982). Several studies have confirmed the importance of class size on student's academic performance in school (see Angrist and Lavy, 1999; Krueger, 1999, 2003; Hoxby, 2000; Dobbelsteen et al., 2002; Rivkin et al., 2005). For example, according to Krueger (2003), the Tennessee Student/Teacher Achievement Ratio experiment (STAR) is one of the most important educational investigations in the class size literature. The STAR project was a longitudinal study in which students and teachers were randomly assigned one of three groups with different class sizes. The design called for students to remain in the same class type for four years after their initial assignment. Over all four years, the sample included 11,600 students from 80 schools. The results showed that students from small class systematically performed better in terms of academic achievement than those from other class. Most of the studies in the literature are consistent with this result. Nonetheless, the existing studies have exclusively focused on the effects of class size in school, and there is no empirical research which examines the impact of class size on student's academic success and course non-completion behaviour at university.

Past research confirms the considerable differences of study area in student dropout or non-retention outcomes (Robst et al., 1998; Rodgers, 2013). Students who study science or engineering may be more likely to drop out than those who study arts or business, mainly due to the degree of difficulty and demands of early papers in these programmes. For example, using administrative data from State University of New York at Binghamton, Robst et al. (1998) found that students in the School of Management are more likely to be retained than students in the School of Arts and Sciences. Our dataset has detailed information about student's enrolled programmes, which enables us to explore the relationship of fields of study and enrolment status (e.g., studying in a double degree programme) on course non-completion behaviour.

Aside from this empirical analysis, many previous studies have made theoretical contributions in modelling student non-completion or dropout behaviour (see Altonji, 1993; Manski, 1989; Light, 1996; Stinebrickner and Stinebrickner, 2012). For example, the student integration model by Tinto (1993) is one of the most comprehensive theoretical models, in which he emphasizes the importance of academic and social integration in predicting retention. Another seminal work is Bean's (1980) student attrition model; this theory extends Tinto's model by incorporating external factors into the student's intention to either stay or leave university. In addition, DesJardins et al. (1999) applied an event history model in examining the temporal dimensions of student departure behaviour. By developing a two-period model, Light and Strayer (2000) investigated whether the "match" between student ability and college quality is an important determinant of college graduation. However, many factors in these theoretical frameworks are hard to quantify, and thus these models cannot be easily converted into a tractable empirical model.

Our research makes three main contributions to the literature. First, our study uses a more comprehensive dataset which enables us to analyse the effects of a wide range of demographic factors, personal characteristics, prior academic histories, and enrolment details on course non-completions. Secondly, all prior work has investigated university student discontinuity behaviour only for programme or degree non-completion; this study is the first time to explore the risk factors associated with paper non-completion. From the policy point of view, investigating this issue for paper non-completion might be more important as it provides opportunities for early intervention services. This research also makes contributions to applying administrative data to study student non-completion behaviour. Using administrative data in this analysis allows us to target high risk groups among all the students; we are also more likely to avoid potential measurement issues and sample selection bias.

Most importantly, past studies almost exclusively focused on the determinants of student degree or programme completion. None of the research has used available information to predict the overall risk status of students. Aside from improving the general understanding of student's discontinuity behaviour, the main objective of this research is to develop predictive risk models (PRM) used to target students at high risk of paper non-completion or programme non-retention before they start their college study. From a policy standpoint, doing so would provide opportunities for policy makers to offer early intervention services.

3. Data

The administrative data used for this analysis were provided by the Department of Strategy and Planning at Auckland University of Technology for this specific project. Data are available on all first-year students who enrolled in Bachelor degree programmes at AUT during the 2009 to 2012 academic years. The full sample contains 18,638 individuals and 101,948 paper observations. Individual observations are utilized to study programme non-retention, while paper observations are used to study paper non-completion as they contain paper-specific information, such as average class size, the overall enrolment in the paper, and the expected study and contact hours for a paper. Variable definitions and descriptive statistics are provided in Table 2 (see the Appendix).

Our dataset contains detailed information on a student's personal characteristics (e.g., ethnicity, country of origin, gender, age, part-time or full-time study, first language status, domestic or international standing), educational background information (e.g., NCEA score, school decile, entrance type), paper characteristics (e.g., average class size, paper size), and student enrolment information (e.g., study area, double degree, number of courses taken by the student).

The two dependent variables used in this study are non-completion of all first-year papers[4] (Non-completion) and programme non-retention in the second year (Non-retention). The dependent variables are dummy variables that are set equal to one if the student had not successfully completed a paper (i.e., received a passing grade), and if the student had not returned to re-enrol at AUT at the beginning of the second year; zero otherwise. The mean non-completion rate is 0.154, indicating that the average paper passing rate is 84.6% in first-year papers in our sample. The mean of non-retention rate is 0.226, indicating that the average retention rate at the beginning of the second year is 77.4% for the 18,638 first-year students in our sample. In this research, the definition of non-retention refers to not returning to AUT by the beginning of the next academic year. Students may leave university permanently or

[4] In this analysis, we do not distinguish between paper non-completion (i.e., individuals who discontinued study prior to the end of the semester and did not complete all assessments) and true fails (i.e., individuals who continued to the end of the semester, completed all assessments, but failed the paper). This is largely because of the government reporting requirements that emphasize paper non-completion outcomes as result of either process.

temporally; possible explanations for dropout include students struggling academically at university, transferring to other institutions, leaving for full-time employment, etc[5].

We include three dummy variables (years 2009, 2010, and 2011) to measure possible differences in student course non-completion behaviour across the years in our sample period. Year 2012 is the omitted category. The means of the three variables indicate that our observations are generally evenly distributed across the four cohorts (see Table 2). We include five dummy variables for a student's self-reported ethnicity (i.e., Asian, European, Māori, Pacific Island, and other ethnicities). The omitted category is the group of students who had not reported their ethnicity. As shown in Table 2, Asian and European students are the main ethnic groups, accounting for 24.2% and 39.2% of first-year students, respectively. Māori and Pacific Island students comprise 9.8% and 11.2%, respectively, of this overall sample of first-year students at AUT.

To examine the impact of country of origin on student's course non-completion behaviour, we include six dummy variables in this analysis (i.e., New Zealand, China, India, Korea, Vietnam, and other countries). New Zealanders (New Zealand citizens or permanent residents) are the single largest group for the first-year students at AUT, constituting nearly 70% of all individuals. Chinese students are the second largest group, with a proportion of 8.6%. It is worth noting that ethnicity and country of origin could be very different in this research due to the fact that New Zealand has historically experienced a substantial inflow of migrants. For

[5] One should be aware that students who struggle at university may be inclined to drop out, but that the best students may also leave university perhaps for more exclusive institutions (Singell, 2004). It should be noted that we have no information in the database on the reasons why individuals may not have returned to study at AUT in the second year.

example, a student who reports “Asian” as his or her ethnicity could also report “New Zealand” as his or her country of origin.

Other personal characteristics variables included being female, enrolling for study part-time, reporting English as the student’s first language, enrolling as a domestic student (see Table 2). We define domestic students as the ones who receive domestic funding status (i.e., government subsidies). The mean age for the 18,638 first-year students is 22.075. The average age might be high compared to other universities or universities from other countries, mainly due to the high portion of part-time students at AUT (29.3%). To account for possible non-linear effects of age in our regression, we create 12 dummy variables for age (i.e., one dummy for ages under 18, eight dummies for ages 18 through 25, and three dummies for the age categories 26 to 30, age 31 to 35, and age 36 to 45). The omitted age category is those 45 years of age and older.

For educational background factors, our dataset has information on the student’s NCEA score, school decile for the last secondary school in which he or she was enrolled, and the entrance type at AUT. As NCEA scores are available for only 44.4% students, we include a dummy variable in our regressions for “Known NCEA” score to indicate whether or not this information was available for a given student. “NCEA Level 3”[6] is the main entrance type for students at AUT, accounting for 36.3% of all first-year students in our sample. The variable “Special Admission”[7] refers to students who typically did not meet all the entrance requirements. Age for students who have used special admission for entry is generally above the average age of all the first-year students (i.e. 22.075). The variables “Internal” and

[6] In this research, “NCEA Level 3” includes NCEA Level 3, Bursary and University Entrance. Bursary and University Entrance are equivalent to NCEA Level 3; they were replaced by NCEA Level 3 in 2004.

[7] In this study, we include discretionary admission in “Special Admission”.

"External" identify students who have held pre-degree certificates or diplomas from AUT or other New Zealand universities and used these qualifications to apply for entry in the Bachelor degree programmes. Including the two variables would be of great importance to test whether students who had studied at university previously would have different course completion outcomes. The variable "Cambridge or IB" refers to those who hold Cambridge or International Bachelaurate programmes at secondary school. These are typically high-achieving students from private high schools, who can use these secondary school qualifications to apply for universities both inside and outside of New Zealand. Note that only 1.4% of first-year students at AUT in our sample had completed either Cambridge or IB programmes.

In terms of paper specific information, the variable "Class Size" indicates the average class size for the paper[8]. "Study Hours" and "Contact Hours" refer to the number of hours over the semester required for study and scheduled class time (combining lectures, tutorials, workshops and labs) in this paper. In addition, it is important to note that first-year Bachelor students typically enrol level 5 papers (83.6%). However, some programmes may require students to study higher level papers (level 6 or 7 papers). Also, students with strong academic background might be able to study higher level courses at their first year, while students with relatively weak background may be required to study some lower level courses (i.e. level 4 papers). We include three dummy variables for papers at level 4, 6, and 7, with level 5 papers as the omitted category.

Furthermore, using individual observations for programme non-retention analysis enables us to generate two interesting variables. These are the total number of papers taken by a student

[8] The average class size is calculated as the total number of students enrolled in a paper divided by the number of the lecture streams. Note that the mean of the variable "Class Size" (38.279) is the average class size of all the paper observations. In Table 2, the means of the variables "Study Hours", "Contact Hours", and "Paper Size" are also calculated as the means of all the paper observations in our sample.

and the portion of higher level of papers (i.e., level 6 or 7) taken by a student in the first year. We also have information on whether a student is enrolled in a double degree programme and whether the study is relegated to a single campus.

Finally, our dataset has detailed information about student's enrolled programme. We include separate dummy variables for 11 largest Bachelor degree programmes. The omitted category includes all of the relatively small Bachelor degree programmes in terms of the number of enrolled students. In addition, it needs to be mentioned that as information is missing for some factors, we always keep the omitted category as the ones who have not reported this information, for example, for ethnicity, country of origin, school decile, and entrance type.

4. Methodology

Estimation of Maximum Likelihood Probit models enables us to explore the determinants of paper non-completion and second year non-retention outcomes. The observed dependent variables are two dummies: paper non-completion (Non-completion) and programme non-retention (Non-retention). The predictor variables include student characteristics, educational background factors, and institutional features. The basic Probit model can be expressed as the following:

$$Y_i^* = \boldsymbol{\beta X_i} + u_i$$

Where Y_i^* is a latent variable associated with a student (or paper). However, we cannot observe Y_i^*, what we observe is the variable Y_i. Y_i is a dummy variable equal to 1 if the student had not returned to university in the second year (or the paper had not been successfully completed in the first year); 0 otherwise. The dummy variable Y_i for the realised outcomes defined by:

$$Y_i = \begin{cases} 1, & if\ Y_i^* > 0 \\ 0, & if\ Y_i^* \leq 0 \end{cases}$$

$\boldsymbol{X_i}$ is a vector of student (or paper) attributes, including all of the predictor variables. $\boldsymbol{\beta}$ is the corresponding vector of parameters to be estimated, and u_i is the random error which is assumed to have a normal distribution. The parameters $\boldsymbol{\beta}$ are estimated by the method of maximum likelihood.

The probability of paper non-completion or programme non-retention can be donated as the following:

$$\mathrm{P_i} = \Pr(Y_i = 1|\boldsymbol{X_i}) = \Pr(Y_i^* > 0) = \Pr(\boldsymbol{\beta X_i} + u_i > 0) = \Pr(u_i > -\boldsymbol{\beta X_i}) = \boldsymbol{\Phi}(\boldsymbol{\beta X_i})$$

Where $\boldsymbol{\Phi}(.)$ is the Cumulative Distribution Function (CDF) of the standard normal distribution.

In this research, we use average marginal effects to describe the effects of a change in the probability of paper non-completion or programme non-retention for a change in one of the elements. This is because Probit model is a non-linear function, and the marginal effects are dependent on the values for all of the other repressors.

$$\frac{\partial \Pr(Y_i = 1|\boldsymbol{X_i})}{\partial \boldsymbol{X_k}} = \beta_k \phi(\boldsymbol{\beta X_i})$$

Where $\phi\ (.)$ is the Probability Distribution Function (PDF) of the standard normal distribution.

More importantly, the Probit models are also utilized to develop our predictive risk models (PRM). The PRM is designed to generate a risk score for every student who had enrolled in a Bachelor degree programme at AUT for the first time. The data are randomly split into two samples (50% of observations for each sample): an “estimation” and “validation” sample, employing the methodology used by Billings et al. (2006 and 2012). The prediction samples will be used to estimate Probit models and the validation samples will be used to assess how well the PRM correctly identifies the actual course non-completion outcomes of students.

Performance of Predictive Risk Models is summarized by reporting the area under the Receiver Operator Characteristic (ROC) curve. The ROC curve characterizes the relationship between sensitivity and specificity. It shows the trade-off between true positives (sensitivity) and false negatives (1-specificity). In this research, the sensitivity is the probability that a student’s course completion outcome is correctly identified by the model. On the contrary, the specificity is defined as the probability that a student’s course completion outcome is incorrectly identified.

The area under the ROC curve quantifies how well the PRM accurately distinguishes students who had completed the courses from those who had not completed the courses. The larger the area under the curve, the better the PRM at assessing risk. A model with 100% area under the ROC curve is said to have perfect fit, while a model with 50% area under the ROC curve is no better than tossing a coin in predicting who is more likely to not complete a paper in the first year or drop out by the start of the second year.

Another approach to assess the predictive power of our PRM is to compare the estimated at-risk scores based on the regression analysis to the actual observed outcomes on paper non-completion or programme non-retention. Specifically, once we have generated risk scores for all the observations in the validation sample, we are able to sort the risk scores by 10 deciles. Then, we could compare the observations with the highest 20% risk scores to their actual outcomes. If the top 20% high risk group only produce 20% of the actual paper non-completion or programme non-retention outcomes in this validation sample, it would be an indicator of the "target ineffectiveness" of this predictive risk tool.

5. Empirical Results

We estimated two separate Probit models for paper non-completion and programme non-retention. As noted above, we randomly selected 50% of our sample for estimation, and used the remaining 50% of sample to validate our PRM[9]. Because the Probit model is nonlinear in parameters, estimated coefficients do not have the usual linear least squares interpretations (i.e., they do not measure the change in the probability of paper non-completion or programme non-retention given a one-unit change in the explanatory variables). Therefore, the marginal effects[10] in this analysis are evaluated at the sample averages of the individual marginal effects. The estimated coefficients, standard errors, and average marginal effects are presented in Table 3 (see the Appendix).

[9] As we select 50% observations from the sample randomly, the estimated coefficients and average marginal effects should be similar to the estimates for the full sample.

[10] The marginal effects could also be calculated at the sample means of the explanatory variables. The Slutsky theorem shows that, for continuous functions in large samples, this technique yields the same answer as the marginal effects evaluated at the sample average of the individual marginal effects (Greene, 2012).

5.1 Paper Non-completion

Our results show that, holding constant measured student, programme and paper characteristics, the probability of paper non-completion is systematically lower in the years 2009, 2010, and 2011, compared to year 2012 as the benchmark (or omitted category). This provides evidence that the paper non-completion probability increased in 2012. Given the lack of any clear time trend in these estimated marginal effects, it would be premature to conclude that these results suggest a systematic increase in paper non-completion rates over time.

Similar to the results for the effects of ethnicity in prior work in New Zealand (Bai and Maloney, 2006), we find race has a significant impact on the probability of paper non-completion. Being Asian or European reduces the risk of paper non-completion, while being Māori or Pacific Island increases the probability of paper non-completion. Specifically, for example, if a paper is taken by a Pacific Island student, holding other factors constant, it increases the probability of non-completion by an average of 6.67 percentage points, compared to a paper observation without reported ethnicity. Māori and Pacific Island students have a substantially higher risk of paper non-completion compared to European and Asian students. These relative comparisons could be made by adding the absolute values of these estimated marginal effects across any two ethnic groups.

The estimated results on country of origin require some explanation. Students from all six reported countries have significantly lower probabilities of paper non-completion, compared to those who had not reported their country of origin. This omitted category appears to be with the highest risk group. Also, students from Vietnam are least likely to not complete the paper; they have nearly half the estimated risk of Korean students.

Consistent with many studies in the literature (Montmarquette et al., 2001, 2007; Mastekaasa and Smeby, 2008), female students have a lower estimated probability of paper non-completion compare to male students. Holding other things constant, being female lowers this probability of paper non-completion by 2.73 percentage points. This effect is statistically significant at better than 1% level. Given the mean probability of paper non-completion in the sample is 15.4%; this gender effect is substantial as it is nearly 18% of this sample mean.

In addition, it is of great importance to note that part-time study has a substantial negative effect on student's paper completion behaviour. The results show that being a part-time student increases the probability of paper non-completion by 15.49 percentage points. This marginal effect is very large in the sense that it is nearly equal to the sample mean of paper non-completion. A reason for this large effect could be that tertiary study requires a student's persistent commitment, while lack of commitment to full-time study or a weaker attachment to university study might have a negative impact on student's academic success.

English as the first language does not appear to be an important factor associated with paper non-completion, as both the variables "Language" and "English" are not statistically significant. Our results suggest that being a domestic student actually increases the probability of paper non-completion by 3.80 percentage points, which is not what we expected. Moreover, the results on the age dummy variables indicate that ages 20 and 21 are the highest risk groups, with the probability of paper non-completion being 7.13 and 7.02 percentage points higher, respectively, when compared to the group of students aged 45 or over. It is interesting that being either younger or older than these two ages reduces the risk of paper non-completion. In fact, being aged 45 or over appears to be the lowest risk age group, since all the included age variables in this regression are estimated to have a positive effect.

Our results recognize the importance of high school performance on a student's paper non-completion outcome. As we expected, good academic records at high school are expected to lower the risk of paper non-completion at university. The results for "NCEA Score" suggest that a 100-point increase in the NCEA score reduces the probability of paper non-completion by nearly 10 percentage points; conditional on this information is available. The combination of the two results on "Known NCEA" and "NCEA Score" suggests that only when the NCEA score exceeds 79 does it begin to reduce the probability of paper non-completion. NCEA results have a more 'continuous' impact on the probability of paper non-completion, and 'turning point' appears to be when this score is more than 79. Also, reporting "Literacy and Numeracy"[11] test increase the risk of paper non-completion by 1.68 percentage points. Students who had not included their literacy and numeracy test could be international students or students who have used other qualifications for entry.

Estimating the influence of school deciles on dropout shows some very interesting results. As we expected, students from the lowest school deciles are more likely to not complete their papers, with school decile 1 as the highest risk school group. However, the negative effect decreases as one moves to a higher school decile only until decile 6. It then begins to increase the risk of paper non-completion. The results indicate a U-shaped marginal effect relationship

[11] Literacy and Numeracy test is used to award the NCEA qualifications. Normally, a student needs to achieve both literacy and numeracy tests at level 2 or above to apply for university admission.

with school decile 6 as the lowest risk group. One possible reason for explaining the positive marginal effects for higher level school deciles could be that many students from decile 9 and 10 schools might have taken the University of Auckland as their first choice as it has historically had a higher domestic ranking than AUT. It is worth noting that all of these estimated effects of school decile are positive and significantly different from zero. These results indicate that those not reporting a school decile (e.g., international students) are in the lowest risk category for paper non-completion.

Consider the impact of entrance status on the probability of paper non-completion. Firstly, those entering through NCEA Level 3 and Special Admissions are statistically more likely to not complete their paper than other non-specified forms of entry. Secondly, students entering through the Cambridge or International Bachelorette programmes have the lowest at-risk status compared to other forms of entry. Thirdly, there is a clear difference between internal (i.e., those who have held a pre-degree certificate or diploma from AUT) and external sources of entry (i.e., those who have held a pre-degree qualification from other universities in New Zealand). Both forms of entry are risk-reducing, even though "-Internal-" is not statistically significant. However, previous external study has nearly five-times the effect in reducing the probability of paper non-completion as previous internal (i.e., AUT) study, and the former result is statistically significant at better than 1% level.

Analysis of the results for paper specific information indicate that study hours and contact hours[12] are not significant determinants of student's paper non-completion behaviour. Some very interesting intuition stems from the results for average class size and overall paper

[12] Study hours and contact hours are the stated hours that students are expected to study and to be in class as indicated in the "Paper Study Guide" at the start of the semester.

enrolment size. The average class size is estimated to have a significant positive impact on paper non-completion (– i.e., being in a bigger class increases the probability of paper non-completion). Nonetheless, the effects are very small in magnitude. An increase in class size by 10 students (nearly one-quarter of the sample mean) is predicted to increase the probability of paper non-completion by 0.1 percentage points (less than 1% of the sample mean on this dependent variable).

Different from “Class Size”, the variable “Paper Size” has a negative impact on a student’s probability of paper non-completion. Increasing the total number of students enrolled for a paper by 100 is predicted to lower the probability of paper non-completion by 0.3 percentage point. Combine the results for “Class Size” and “Paper Size”, we find that small class is beneficial for students to complete their paper, and being enrolled in a big course also helps to reduce the risk of paper non-completion. This is because big papers at AUT are typically the first-year compulsory courses that may be generally less difficult in terms of academic content and expectations. In addition, the complexity of the paper appears to have a significant influence on paper non-completion, being a lower level paper (i.e. level 4), on average, decreases this risk by 5.68 percentage points, compared to level 5 paper.

Finally, our results suggest that study area could play an important role in a student’s paper non-completion behaviour. For example, compare to other small programmes, being enrolled in Bachelor of Arts, on average, increases the probability of paper non-completion by 1.72 percentage points. Bachelor of Mathematical Science is the programme with highest risk perhaps due to the difficulty of the course content, while Bachelor of Education is estimated to be the lowest risk programme. From our results, we could characterize the high risk programmes are Bachelor of Arts, Bachelor of Engineering Technology, and Bachelor of Mathematical Science; the low risk programmes are Bachelor of Education, Bachelor of Design, and Bachelor of Communication Studies.

5.2 Programme Non-retention

Estimation for programme non-retention indicates that being enrolled in the 2010 and 2011 cohorts actually increases the probability of dropout by the start of the second year at university by 2.74 and 2.96 percentage points, respectively, compared to the year 2012. Most of the ethnicity variables have the expected signs, but only "Māori" is statistically significant. Specifically, Māori students have a probability of non-retention in the second year that is 5.85 percentage points higher than students without a reported ethnicity. If we pull together the results on ethnicity in both regressions, it shows that Pacific Island students are relatively more likely to not complete their papers in the first year, while Māori students are more likely to not return to the university in the second year.

Consider the results for the country of origin variables. As shown in Table 3, the marginal effects frequently have the same sign as in the paper non-completion analysis, but have larger effects in magnitude on the probability of non-retention. Those who had not reported their country of origin are still the most at-risk group. Vietnamese students are least likely to drop out compared to other groups, and Korean students are the highest at-risk group compared to other students reporting their country of origin (i.e., nearly four-times higher in the risk of dropping out compared to Vietnamese students).

Moreover, our results indicate that the decision to drop out depends significantly on part-time or full-time status; being part-time study increases the probability of programme non-retention by 15.80 percentage points. This result is in accordance with Montmarquette et al. (2001). One explanation for this might be that the opportunity costs of schooling are higher for part-time student. In addition, female students are more likely to both complete their papers in the first year and be retained in the second year, compared to male students. It is interesting to note that compared to the paper non-completion analysis, being a domestic student has even a larger positive impact on student's dropout behaviour. Domestic students have an estimated probability of not being retained in the second year that is 6.64 percentage points higher than international students. Our empirical results provide some of the first direct evidence that domestic students are less likely to return to university in the second year compared to international students. Age variables appear not to play an important role in affecting student's dropout decision since none of the estimated coefficients are statistically significant.

The results for "NCEA Score" indicate that having performed well at high school improves retention. Students who are academically stronger at high school are less likely to drop out by the start of the second year. Specifically, a 100-point increase in the NCEA score (conditional on this information being available), on average, reduces the probability of programme non-retention by 6 percentage points. This suggests the importance of improving academic preparation at high school for both paper completion rates in the first year and retention in the second year at university.

Similar to paper non-completion analysis, including literacy and numeracy test is predicted to increase the risk of non-retention by 3.74 percentage points. In addition, unlike a student's paper non-completion behaviour, the results for school deciles show very different outcomes. As we expected, school decile 1 is the highest risk group of university non-retention, students from this lowest school decile are predicted to increase the probability of non-retention by 5.30 percentage points. Different from the results of the top school decile in paper non-completion analysis, the marginal effects of school decile 10 on non-retention are estimated to be negative. It is also interesting to note that schools from decile 4 and above are typically more likely to decrease the risk of programme non-retention, while the marginal effects for schools under decile 4 are generally positive. Thus, there is some empirical evidence for the threshold level for high-risk of dropping out at university being at school decile 4.

Consider the results for the entrance type variables. Similar to paper non-completion behaviour, being students who had used Cambridge or International Baccalaureate qualifications for entry decreases the risk of dropout by 9.48 percentage points. In addition, the results for “Internal” and “External” suggest that having studied at the same university or other universities by the time of matriculation does help to reduce the risk of dropout. This is might because these students have lower costs in integrating with or acclimating to university (e.g., see Bean, 1980; Tinto, 1993). They are generally more familiar with university life compare to those who are new to university. This work provides the first evidence about the importance of pre-experience at university on student’s dropout behaviour.

In terms of other academic factors, having enrolled in a double degree programme is estimated to have a substantial impact on reducing the risk of dropout; it decreases the probability by 13.11 percentage points. This is because students enrolled for a double degree might process unobservable factors that also reduce the risk of non-retention, such as intelligence, perseverance and other personality traits that result in them continuing their university studies beyond the first year. It is also important to note that having a higher proportion of high level papers appears to have a negative impact on programme non-retention. Increasing the proportion of level 6 or 7 papers in the first year by 10% is predicted to reduce the risk of dropping out in the second year by an average of 2.08 percentage points. In addition, whether or not studying at the same campus appears to play no role in a student's dropout or paper non-completion probabilities.

Finally, consider the effects of study area on student's dropout behaviour. Although some estimates are not statistically significant, our results provide some interesting intuition. For example, compared to other small Bachelor degree programmes, being enrolled in Bachelor of Arts increases the probability of programme non-retention by 8.73 percentage points. This risk is approximately 12 percentage points higher than that of Bachelor of Business, which is the largest Bachelor degree programme at AUT. Moreover, Bachelor of Education and Bachelor of Communication Studies are at low risk of dropping out; while Bachelor of Sports and Recreation is at high risk of non-retention at university in the second year.

5.3 Predictive Power of Our PRM

As noted above, we randomly selected 50% of our sample for estimation purposes, and used the remaining 50% of the sample for assessing the performance of our PRM. We examined the predictive power of our PRM by reporting the area under Receiver Operator Characteristic (ROC) curves for paper non-completion and programme non-retention, respectively. These results are shown in Figures 1 and 2 in the Appendix.

The Receiver Operator Characteristic (ROC) curves graphically illustrate the trade-offs between sensitivity and 1 minus specificity at all possible thresholds. The area under the curve for paper non-completion is 0.7553, indicating that there is a 75.53% probability that a randomly selected paper observation with a non-completion outcome will receive a higher risk score by our PRM than a randomly selected paper observation with a completion outcome. This is an indicator of the ‘target effectiveness’ of this predictive risk tool, compared to the results in the literature (see Billings et al., 2006, 2012; Vaithianathan et al., 2013). Similar interpretations can be given for the non-retention analysis in the second year with the area under ROC curve at 0.7125.

Another approach to examine the effectiveness of our PRM is to test its predictive power in identifying the high risk observations (-see Table 4 in the Appendix). Suppose, for example, for paper non-completion analysis, we wanted to intervene (i.e., provide specific services) to those with the highest 20% of risk scores. We know from our validation data that this high-risk group would produce 47.57% of actual paper non-completions. Suppose that we more narrowly concentrated limited resources on those with the highest 10% of risk scores. This would account for 29.25% of paper non-completions. For such an amount of predictive factors, this is a relatively good result as a predictive risk tool. Similar interpretations could be provided for Programme our non-retention analysis (i.e., the highest 20% and 10% of risk scores would produce 40.91% and 23.33% of actual non-retentions, respectively)[13].

Overall, the general profile of a high-risk individual from this analysis would be a male Māori or Pacific Island domestic student, enrolled part-time in high-risk programmes, not reporting a country of origin, with a literacy and numeracy test and a low NCEA score, with no previous university study, not entering through a Cambridge or International Bachelorette programme, and coming from a decile 1 school.

5.4 One Experiment

It would be of great interest if we include paper non-completion rates as one predictor variable when developing PRM for programme non-retention analysis. This analysis, for example, might be useful to programme administrators using a PRM approach to identifying students most at-risk of non-retention *at the end of* the first year of university study. To do so would allow us to estimate the relationship between paper non-completion and programme non-retention. The third column of Table 4 presents the results of this experiment. These findings

[13] Note that AUT had previously used a simple risk analysis tool that relied on limited information to predict those most at risk of academic difficulties in the first year. This was done by arbitrarily setting weights to a selected set of variables. According to this tool, the top two high risk categories would account for 11.78% of observations and 23.51% of actual non-completions. The top three categories would account for 25.27% of observations and 39.11% of the non-completions. We can compare this ad hoc procedure to our PRM. By targeting just the top 10% of risk scores, our PRM model captures 29.25% of non-completions. Targeting the top 20% captures 47.57% of non-completions. Our systematic approach has a much better “hit rate” than this previous subjective procedure.

suggest that it would improve the predictive power of PRM dramatically by increasing the area under ROC curve from 0.7125 to 0.8694. Also, the top risk decile (i.e. those with the highest 10% of risk scores) would produce 40.63% of non-retention outcomes at university at the beginning of the second year. The average marginal effect of paper non-completion is estimated to be 47.45% with significance level better than 1%, indicating that increasing the paper non-completion by 50% would increase the probability of programme non-retention by 23.73 percentage points. The estimation results support the notion that paper non-completion rates in the first year would be one of the most important factors determining student dropout behaviour in subsequent years.

However, although it was always possible to include paper non-completion as one predictor variable when developing the PRM for non-retention, doing so would be less useful from the predictive point of view at the outset of a student's time at university. This is because our PRM is developed to predict the risk of course non-completion for the first-year students at the beginning of their first semester, thereby we could identify the students at high risk of such events and we have enough time to provide intervention services. Nonetheless, once we have observed the paper non-completion outcomes, it would be the end of the first year and it is would leave very scarce time (i.e., only summer break) for the university to deliver possible interventions.

6. Conclusion

This study empirically examines the determinants associated with a student's paper non-completion outcomes during the first year of study and his or her non-retention behaviour in the second year at university, using large-scale administrative database from the Auckland University of Technology. More importantly, we develop a Predictive Risk Model (PRM) using maximum likelihood Probit analysis to identify those students most at-risk of paper non-completion and university non-retention. Our PRM should be of particular interest to both government policy makers and university administrators as it provides opportunities to target and deliver early intervention services to avoid these adverse university outcomes.

Our findings show that many factors from individual characteristics, to educational and institutional characteristics play a significant role in determining a student's probability of paper non-completion in the first year and non-retention in the second year. For example, part-time study is estimated to have a substantial impact in raising a student's probabilities of both paper non-completion and subsequent dropout behaviour. For ethnicity, Pacific Island students are relatively more likely to not successfully complete their first-year papers, while Māori students are relatively more likely to not be retained for their second year of study at AUT. Asian and European students are relatively less likely to be at high risk of both negative outcomes. Also, female students are more likely to complete their papers and return to the university in the second year, compared to male students. It is interesting to note that being a domestic student is predicted to increase the risk of both paper non-completion and programme non-retention.

Educational background factors appear to play an important role in explaining student's non-completion and discontinuity behaviour. Students with better pre-university academic records are less likely to be at high risk of both negative outcomes. Those from lower school deciles tend to be more likely to not complete their papers in the first year and drop out by the second year. Previous study experience at universities helps to reduce the probability of paper non-completion and university non-retention. Our research provides the first direct evidence that highlights the importance of the previous experience at university in this regard.

Institutional characteristics, including average class size, paper size, and the complexity of the course, would also be the factors associated with paper non-completion and attrition. Moreover, there are substantial differences of study area in student's non-completion and non-retention. Enrolling in programmes, such as the Bachelor of Arts, Bachelor of Engineering Technology, and Bachelor of Mathematical Science, is predicted to increase the probability of paper non-completion and non-retention. While some programmes, for instance, Bachelor of Education and Bachelor of Design, are estimated to be at low risk of both detrimental outcomes.

Our data were randomly spilt into an "estimation" and "validation" sample. We used 50% of our sample to develop our PRM and used the remaining half of the sample to assess the predictive power of our PRM. The areas under ROC curve are 0.7553 and 0.7125, respectively, indicating the overall "target effectiveness" of the tool. The top risk decile (10%) could produce nearly one quarterly of actual paper non-completion or programme non-retention. The top 20% of at-risk scores according to our PRM analysis would account for nearly 48% and 41% of actual paper non-completion and university non-retention outcomes, respectively. The top 10% of at-risk scores would still account for 29% and 23% of actual non-completion and non-retention outcomes, respectively.

Our research provides good evidence that Predictive Risk Modelling (PRM) can be utilized in predicting poor academic outcomes at university. Further studies along this line could benefit from using many financial factors, such as financial aid, student scholarships or loans (e.g., see Stampen and Cabrera, 1988; McPherson and Schapiro, 1991; Card and Lemieux, 2000; DesJardins et al., 2002; Singell, 2004; Kerkvliet and Nowell, 2005). According to Stratton et al. (2008), financial aid positively affects persistence by reducing the costs associated with completing courses. Furthermore, it would of particular interest to include more detailed academic information from schools, such as specific grades and exam results, and course of study at high school. Finally, it would also be useful to include family factors in further research, such as family income and parental education, since many studies have found such factors play a prominent role in a student's academic success at university (e.g., see Kerkvliet and Nowell, 2005; Ishitani, 2006; Montmarquette et al., 2007; Stratton et al., 2008; Belloc et al., 2010). Despite the above data issues, our use of available administrative data has provided valuable insight into understanding the factors associated with course non-completion, and more importantly, have provided a potential platform for the future use of administrative data for targeting early intervention services to students most at-risk of poor academic outcomes at university.

References

Altonji, J. G. (1993). The demand for and return to education when education outcomes are uncertain. *Journal of Labor Economics, 11*(1), 48-83.

Angrist, J., and Lavy, V. (1999). Using Maimonides's rule to estimate the effect of class size on children's academic achievement. *Quarterly Journal of Economics*, 114, 533-575.

Bai, J., and Maloney, T. (2006). Ethnicity and academic success at university. *New Zealand Economic papers, 40*(2), 181-213.

Bean, J. P. (1980). Dropouts and turnover. The synthesis and test of a causal model of student attrition. *Research in Higher Education*, 12(2), 155-187.

Belloc, F., Maruotti, A., and Petrella, L. (2010). University drop-out: an Italian experience. *Higher Education*, 60, 127-138.

Betts, J. R., and Morell, D. (1999). The determinants of undergraduate grade point average: The relative importance of family background, high school resources, and peer group effects. *Journal of Human Resources, 34*(2), 268-293.

Billings, J., Blunt, I., Steventon, A., Georghiou, T., Lewis, G., and Bardsley, M. (2012). Development of a predictive model to identify inpatients at risk of re-admission within 30 days of discharge (PARR-30). *BMJ Open*, 00:e001667.

Billings, J., Dixon, J., Mijanovich, T., and Wennberg, D. (2006). Case findings for patients at risk of readmission to hospital: development of algorithm to identify high risk patients. *BMJ*, doi: 10.1136/bmj.38870.657917.

Card, D., and Lemieux, T. (2000). Droupout and enrolment trends in the post-war period: What went wrong in the 1970s? NBER Working Paper No.7658.

Cohn, E., Cohn, S., Balch, D. C., and Bradley, J. (2004). Determinants of undergraduate GPAs: SAT scores, high-school GPA and high school rank. *Economics of Education Review*, 23, 577-586.

Cyrenne, P., and Chan, A. (2012). High school grades and university performance: A case study. *Economics of Education Review*, 31, 524-542.

DesJardins, S. L., Ahlburg, D. A., and McCall, B. P. (1999). An event history model of student departure. *Economics of Education Review*, 18, 375-390.

DesJardins, S. L., Ahlburg, D. A., and McCall, B. P. (2002). Simulating the longitudinal effects of changes in financial aid on student departure from college. *Journal of Human Resources, 37*(3), 653-679.

Dobbelsteen, S., Levin, J., and Oosterbeek, H. (2002). The causal effect of class size on scholastic achievement: Distinguishing the pure class size effect from the effect of changes in class composition. *Oxford Bulletin of Economics and Statistics*, 64, 17-38.

Ficano, C. C. (2012). Peer effects in college academic outcomes – gender matters! *Economics of Education Review*, 31, 1102-1115.

Grayson, J. P. (1998). Racial origin and student retention in a Canadian University. *Higher Education*, 36, 323-352.

Greene, W. H. (2012). *Econometric analysis*, 7th edition. Prentice Hall.

Grubb, W. N. (1989). Dropouts, spells of time, and credits in postsecondary education: Evidence from longitudinal surveys. *Economics of Education Review*, 8, 49-67.

Hartog, J., Pfann, G., and Ridder, G. (1989). (Non-)Graduation and the earning function: An inquiry on self-selection. *European Economic Review*, 33, 1371-1395.

Hoxby, C. M. (2000). The effects of class size on student achievement: New evidence from population variation. *Quarterly Journal of Economics*, 115, 1239-1285.

Ishitani, T. T. (2006). Studying attrition and degree completion behaviour among first-generation college students in the United States. *Journal of Higher Education, 77*(5), 861-885.

Kerkvliet, J., and Nowell, C. (2005). Does one size fit all? University differences in the influence of wages, financial aid, and integration on student retention. *Economics of Education Review*, 24, 85-95.

Krueger, A. B. (1999). Experimental estimates of education production functions. *Quarterly Journal of Economics*, 114(2), 497-532.

Krueger, A. B. (2003). Economic considerations and class size. *Economic Journal*, 113, F34-F63.

Light, A. (1996). Hazard model estimates of the decision to reenrol in school. *Labor Economics*, 2, 381-406.

Light, A., and Strayer, W. (2000). Determinants of college completion: School quality or student ability? *Journal of Human Resources, 35*(2), 299-332.

Manski, C. F. (1989). Schooling as experimentation: A reappraisal of the postsecondary dropout phenomenon. *Economics of Education Review, 8*(4), 305-312.

Mastekaasa, A., and Smeby, J. C. (2008). Educational choice and persistence in male- and female-dominated fields. *Higher Education*, 55, 189-202.

McPherson, M. S., and Schapiro, M. O. (1991). Does student aid affect college enrolment? New evidence on a persistence controversy. *American Economic Review, 81*(1), 309-318.

Montmarquette, C., Mahseredjian, S., and Houle, R. (2001). The determinants of university dropouts: a bivariate probability model with sample selection. *Economics of Education Review*, 20, 475-484.

Montmarquette, C., Viennot-Briot, N., and Dagenais, M. (2007). Dropout, school performance, and working while in school. *Review of Economics and Statistics, 89*(4), 752-760.

New Zealand Ministry of Education. (2004). *Retention, completion and progression in tertiary education 2003.* Wellington: Ministry of Education.

Ost, B. (2010). The role of peers and grades in determining major persistence in the sciences. *Economics of Education Review, 29*(6), 923-934.

Rivkin, S. G., Hanushek, E. A., and Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417-458.

Rodgers, T. (2013). Should high non-completion rates amongst ethnic minority students be seen as an ethnicity issue? Evidence from a case study of a student cohort from a British University. *Higher Education, 66*(5), 535-550.

Robst, J., Keil, J., and Russo, D. (1998). The effect of gender composition of faculty on student retention. *Economics of Education Review, 17*(4), 429-439.

Singell, L. D. (2004). Come and stay a while: does financial aid effect retention conditioned on enrolment at a large public university? *Economics of Education Review*, 23, 459-471.

Stampen, J. O., and Cabrera, A. F. (1988). The targeting and packaging of student aid and its effect on attrition. *Economics of Education Review, 7*(1), 29-46.

Stinebrickner, T., and Stinebrickner, R. (2012). Learning about academic ability and the droupout decision. *Journal of Labor economics, 30*(4), 707-748.

Stratton, L. S., O'Toole, D. M., and Wetzel, J. N. (2008). A multinomial logit model of college stopout and dropout behaviour. *Economics of Education Review*, 27, 319-331.

Tinto, V. (1982). Limits of theory and practice in student attrition. *Journal of Higher Education*, 53, 687-700.

Tinto, V. (1993). *Leaving college: Rethinking the causes and cures of student attrition*, *2nd Edition*. Chicago: Chicago University Press.

Vaithianathan, R., Maloney, T., Putnam-Hornstein, E., and Jiang, N. (2013). Using predictive modelling to identify children in the public benefit system at high risk of substantiated maltreatment. *American Journal of Preventive Medicine, 45*(3), 354-359.

Venti, S. F., and Wise, D. A. (1982). Test scores, educational opportunities, and individual choice. *Journal of Public Economics*, 18, 35-63.

Wetzel, J. N., O'Toole, D. M., and Peterson, S. (1999). Factors affecting student retention probabilities: A case study. *Journal of Economics and Finance, 23*(1), 45-55.

Appendix:

Table 2		
Descriptive statistics and variable definitions		
Variable	Definition	Mean (std. deviation)
Dependent variables		
Non-completion	1 if paper is not successfully completed	0.154 (0.361)
Non-retention	1 if student does not return to AUT	0.226 (0.418)
Year of cohort		
Year 2009	1 if student enrols in the year 2009	0.225 (0.478)
Year 2010	1 if student enrols in the year 2010	0.253 (0.435)
Year 2011	1 if student enrols in the year 2011	0.240 (0.427)
Year 2012	the omitted category for student who enrols in the year 2012	0.281 (0.450)
Ethnicity		
Asian	1 if student's ethnicity is Asia	0.242 (0.428)
European	1 if student's ethnicity is European	0.392 (0.488)
Māori	1 if student's ethnicity is Māori	0.098 (0.297)
Pacific	1 if student's ethnicity is Pacific Island	0.112 (0.316)
Others	1 if student's ethnicity is none of above	0.080 (0.272)
Unknown	the omitted category for student who had not reported his or her ethnicity	0.076 (0.265)
Country of origin		
NZ	1 if student's origin is New Zealand	0.695 (0.460)
China	1 if student's origin is China	0.086 (0.280)
India	1 if student's origin is India	0.016 (0.127)
Korea	1 if student's origin is Korea	0.022 (0.147)
Vietnam	1 if student's origin is Vietnam	0.013 (0.113)
Others	1 if student's origin is none of above	0.155 (0.362)
Unknown	the omitted category for student who had not reported his or her country of origin	0.013 (0.111)
Personal characteristics		
Female	1 if student is female	0.601 (0.490)
Part-time Enrol	1 if student is enrolled as part-time	0.293 (0.455)

Language	1 if student's first language is reported	0.578 (0.494)
English	1 if student's first language is English (conditional on one's first language is reported)	0.707 (0.455)
Domestic	1 if student receives domestic funding	0.879 (0.326)
Age	12 dummy variables, i.e. under age 18, eight dummies for ages 18 through 25, three dummies for the age 26 to 30, age 31 to 35, and age 36 to 45 1 if student’s age is in this category	22.075 (6.322)
Pre-academic information		
Known NCEA	1 if student's NCEA score is reported	0.444 (0.497)
NCEA Score	student's NCEA score (conditional on one's score is reported)	155.107 (62.860)
L&N Included	1 if student has taken literacy and numeracy test in high school	0.238 (0.426)
School Decile	10 dummy variables, i.e. school decile 1 to 10 1 if student's school decile is in this category	6.846 (2.812)
Entrance type		
NCEA Level 3	1 if student's entrance type is NCEA level 3 (including Bursary entrance and University entrance)	0.363 (0.481)
Special Admission	1 if student's entrance type is special admission (including discretionary admission)	0.130 (0.336)
Internal	1 if student holds a pre-degree from AUT University	0.089 (0.285)
External	1 if student holds an equivalent pre-degree from other New Zealand universities	0.150 (0.358)
Cambridge/IB	1 if student's entrance type is Cambridge or International Bachelaureate	0.014 (0.117)
Unknown	the omitted category for student who had not reported his or her entrance type	0.253 (0.435)
Paper information		
Study Hours	study hours for the paper	180.539 (62.964)
Known Con	1 if contact hours for the paper are reported	0.844 (0.362)
Contact Hours	contact hours for the paper (conditional on the contact hours are reported)	75.980 (32.234)
Class Size	average class size of the paper	38.279(28.932)
Paper Size	total number of students enrolled in the paper	562.194 (535.464)
Internet	1 if paper is supported by Internet	0.588 (0.492)
Level 4	1 if paper is level 4 (in terms of complexity)	0.005 (0.067)
Level 5	the omitted category as the benchmark	0.836 (0.371)
Level 6	1 if paper is level 6	0.156 (0.363)
Level 7	1 if paper is level 7	0.004 (0.059)
Individual academic information		
Course Number	number of courses taken by a student per year	5.243 (2.301)
High Level	proportion of level 6 or 7 courses taken by a student	0.138 (0.202)

Double Degree	1 if student's programme is double-degree	0.008 (0.088)
One Campus	1 if student's study is at one campus only	0.913 (0.283)
Programmes		
BA	1 if student enrols in Bachelor of Arts	0.078 (0.268)
BBus	1 if student enrols in Bachelor of Business	0.282 (0.450)
BCIS	1 if student enrols in Bachelor of Computer Information Science	0.049 (0.216)
BCS	1 if student enrols in Bachelor of Communication Studies	0.068 (0.250)
BDe	1 if student enrols in Bachelor of Design	0.074 (0.262)
BEdu	1 if student enrols in Bachelor of Education	0.040 (0.197)
BEngT	1 if student enrols in Bachelor of Engineering Technology	0.029 (0.168)
BHS	1 if student enrols in Bachelor of Health Science	0.195 (0.396)
BIHM	1 if student enrols in Bachelor of International Hospitality Management	0.043 (0.204)
BMS	1 if student enrols in Bachelor of Mathematical Science	0.006 (0.079)
BSR	1 if student enrols in Bachelor of Sports and Recreation	0.060 (0.237)
Others	the omitted category for other small Bachelor programmes	0.075 (0.291)

Table 3

Probit model for parameter estimates and average marginal effects

	Paper non-completion			Programme non-retention		
Variable	parameter	std. error	dy/dx	parameter	std. error	dy/dx
Constant	-0.6693***	0.1148	-	-0.1374	0.2156	-
Year of cohort						
Year2009	-0.1069***	0.0217	-2.20%	-0.0134	0.0447	-0.36%
Year2010	-0.0752***	0.0198	-1.54%	0.1030**	0.0424	2.74%
Year2011	-0.1725***	0.0202	-3.54%	0.1112***	0.0428	2.96%
Ethnicity						
Asian	-0.1267***	0.0442	-2.60%	-0.1079	0.0901	-2.87%
European	-0.1830***	0.0483	-3.76%	0.0164	0.0984	0.44%
Māori	0.1666***	0.0515	3.42%	0.2200**	0.1064	5.85%
Pacific	0.3247***	0.0501	6.67%	0.0967	0.1040	2.58%
Others	-0.0417	0.0511	-0.86%	-0.0525	0.1048	-1.40%
Country of origin						
NZ	-0.4828***	0.0630	-9.91%	-0.7431***	0.1258	-19.78%
China	-0.4488***	0.0714	-9.21%	-0.8090***	0.1411	-21.53%
India	-0.4670***	0.0865	-9.59%	-0.7482***	0.1777	-19.91%
Korea	-0.3356***	0.0809	-6.89%	-0.4329***	0.1616	-11.52%
Vietnam	-0.6662***	0.1081	-13.68%	-1.5075***	0.2574	-40.12%
Others	-0.4833***	0.0650	-9.92%	-0.8868***	0.1299	-23.60%
Personal characteristics						
Female	-0.1328***	0.0165	-2.73%	-0.0583*	0.0343	-1.55%
Part-time Enrol	0.7546***	0.0179	15.49%	0.5935***	0.0466	15.80%
Language	0.0302	0.0250	0.62%	0.0304	0.0523	0.81%
English	0.0057	0.0269	0.12%	-0.0144	0.0568	-0.38%
Domestic	0.1852***	0.0434	3.80%	0.2495***	0.0880	6.64%
Under 18	0.1244	0.1098	2.55%	-0.2035	0.2180	-5.42%
Age 18	0.1854***	0.0645	3.81%	-0.1700	0.1243	-4.53%
Age 19	0.2233***	0.0646	4.58%	-0.0735	0.1245	-1.96%
Age 20	0.3473***	0.0649	7.13%	-0.0037	0.1250	-0.10%

Age 21	0.3418***	0.0658	7.02%	0.0469	0.1272	1.25%
Age 22	0.2411***	0.0679	4.95%	-0.0147	0.1323	-0.39%
Age 23	0.2555***	0.0695	5.25%	-0.082	0.1353	-2.18%
Age 24	0.1512**	0.0730	3.10%	-0.0653	0.1421	-1.74%
Age 25	0.1427*	0.0758	2.93%	0.0822	0.1439	2.19%
Age 26 to 30	0.0764	0.0665	1.57%	0.0286	0.1251	0.76%
Age 31 to 35	0.0264	0.0730	0.54%	-0.1324	0.1364	-3.52%
Age 36 to 45	0.0838	0.0720	1.72%	-0.119	0.1365	-3.17%
Pre-academic information						
Known NCEA	0.3703***	0.0344	7.60%	0.1991***	0.0741	5.30%
NCEA Score	-0.0047***	0.0002	-0.10%	-0.0024***	0.0005	-0.06%
L&N Included	0.0819***	0.0259	1.68%	0.1404***	0.0473	3.74%
School Decile 1	0.4710***	0.0423	9.67%	0.084	0.0938	2.23%
School Decile 2	0.2370***	0.0419	4.87%	-0.1128	0.0899	-3.00%
School Decile 3	0.1823***	0.0374	3.74%	0.0225	0.0804	0.60%
School Decile 4	0.1674***	0.0326	3.44%	-0.1706**	0.0696	-4.54%
School Decile 5	0.0865**	0.0391	1.78%	-0.0602	0.0812	-1.60%
School Decile 6	0.0740**	0.0374	1.52%	-0.1361*	0.0778	-3.62%
School Decile 7	0.0886***	0.0340	1.82%	-0.1345*	0.0718	-3.58%
School Decile 8	0.1264***	0.0349	2.60%	0.0019	0.0724	0.05%
School Decile 9	0.1415***	0.0323	2.91%	-0.063	0.0664	-1.68%
School Decile 10	0.1607***	0.0289	3.30%	-0.1293**	0.0601	-3.44%
Entrance type						
NCEA Level 3	0.1752***	0.0363	3.60%	-0.0128	0.0768	-0.34%
Spec. Admission	0.0752***	0.0270	1.54%	-0.0827	0.0559	-2.20%
Internal	-0.0312	0.0310	-0.64%	-0.2004***	0.0653	-5.33%
External	-0.1542***	0.0274	-3.17%	-0.1752***	0.0556	-4.66%
Cambridge/IB	-0.3138***	0.0703	-6.44%	-0.3563**	0.1517	-9.48%
Paper information						
Study Hours	0.0030	0.0265	0.06%	-	-	-
Known Con	-0.0702**	0.0346	-1.44%	-	-	-
Contact Hours	-0.0101	0.0480	-0.21%	-	-	-
Class Size	0.0007**	0.0003	0.01%	-	-	-
Paper Size	-0.0147***	0.0027	-0.30%			
Internet	0.0187	0.0187	0.38%	-	-	-
Level 4	-0.2765***	0.1049	-5.68%	-	-	-
Level 6	-0.0218	0.0228	-0.45%	-	-	-

Level 7	-0.1097	0.1164	-2.25%	-	-	-
Individual academic information						
Course Number	-	-	-	-0.001	0.0119	-0.03%
High Level	-	-	-	-0.7818***	0.0963	-20.81%
Double Degree	-0.1524	0.0936	-3.13%	-0.4927**	0.2423	-13.11%
One Campus	-0.0075	0.0247	-0.15%	0.0446	0.0559	1.19%
Programmes						
BA	0.0837***	0.0316	1.72%	0.3280***	0.0748	8.73%
BBus	-0.1791***	0.0469	-3.68%	-0.1242**	0.0719	-3.30%
BCIS	0.0167	0.0372	0.34%	-0.0737	0.0868	-1.96%
BCS	-0.2971***	0.0410	-6.10%	-0.1786**	0.0887	-4.75%
BDe	-0.3306***	0.0397	-6.79%	-0.1221	0.0818	-3.25%
BEdu	-0.5306***	0.0456	-10.89%	-0.2231**	0.0948	-5.94%
BEngT	0.0896*	0.0462	1.84%	-0.1237	0.1064	-3.29%
BHS	-0.3598***	0.0328	-7.39%	-0.033	0.0642	-0.88%
BIHM	-0.2196***	0.0407	-4.51%	-0.0913	0.0930	-2.43%
BMS	0.1758**	0.0727	3.61%	0.1497	0.1930	3.98%
BSR	-0.1145***	0.0366	-2.35%	0.1480*	0.0795	3.94%
Pseudo R2		0.1339			0.1063	
Log-likelihood		-18985.6			-4417.61	
Under ROC curve		0.7553			0.7125	
Number of observations		50932			9301	
Note: * indicates significance at the 1% level, ** indicates significance at the 5% level, * indicates significance at the 10% level.**						

Table 4

Predictive power of PRM

	paper non-completion	programme non-retention	programme non-retention*
Top 1 decile (10%)	29.25%	23.33%	40.63%
Top 2 deciles (20%)	47.57%	40.91%	64.88%
Under ROC curve	0.7553	0.7125	0.8694
Number of observations	50932	9301	9346

Note: a. The model for programme non-retention* includes paper non-completion as one independent variable.
b. Paper non-completion as an independent variable in the model for programme non-retention* is statistically significant at 1% level, with average marginal effect at 47.45%.

ROC curve for paper non-completion

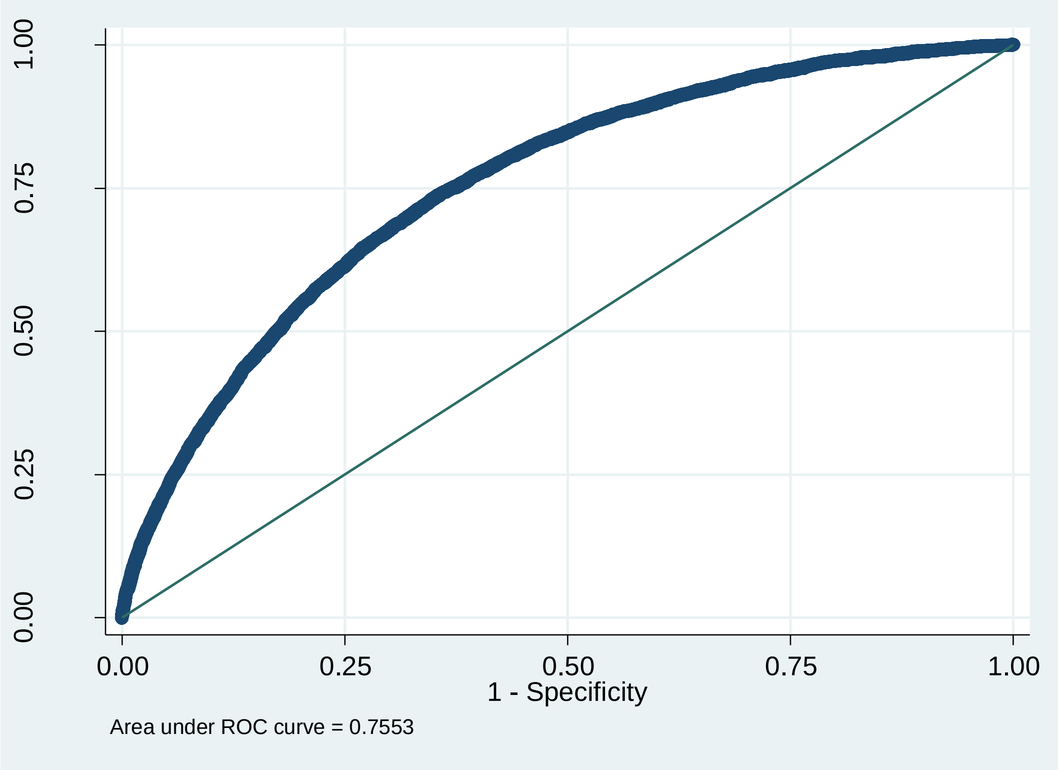

Figure 1

ROC curve for programme non-retention

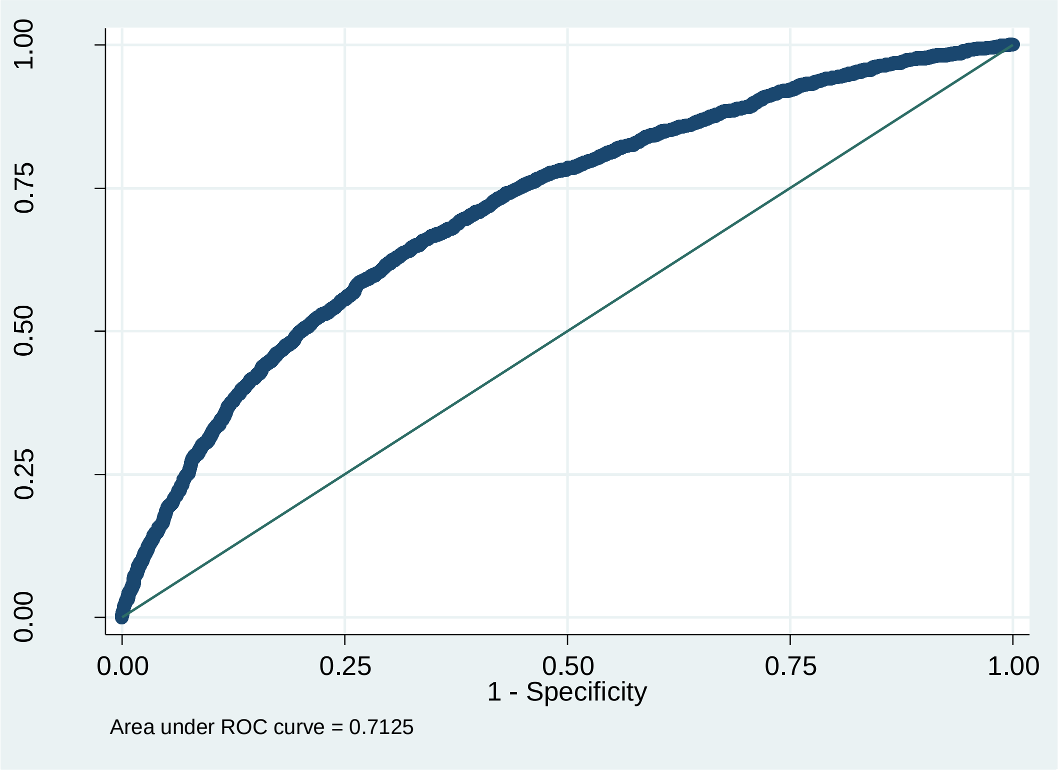

Figure 2