High Educ (2015) 70:127–149
DOI 10.1007/s10734-014-9829-7

Using predictive modelling to identify students at risk of poor university outcomes

Pengfei Jia · Tim Maloney

Published online: 26 November 2014

Abstract Predictive modelling is used to identify students at risk of failing their first-year courses and not returning to university in the second year. Our aim is twofold. Firstly, we want to understand the factors that lead to poor first-year experiences at university. Secondly, we want to develop simple, low-cost tools that would allow universities to identify and intervene on vulnerable students when they first arrive on campus. This is why we base our analysis on administrative data routinely collected as part of the enrollment process from a New Zealand university. We assess the 'target effectiveness' of our model from a number of perspectives. This approach is found to be substantially more predictive than a previously developed risk tool at this university. For example, observations from validation samples in the top decile of risk scores account for nearly 28 % of first-year course non-completions and 22 % of second-year student non-retentions at this university.

Keywords Educational finance and efficiency · Resource allocation · Predictive risk modelling · University dropout behavior · New Zealand

JEL Classification I21 · I22 · I28

Introduction

Increasingly, poor university outcomes are a concern to students, institutions, and public funding bodies. This may be a by-product of rapidly rising university participation rates over recent decades in many countries.[1] Course non-completion and dropout rates may be

[1] For example, previous studies have analyzed the reasons behind rising university dropout rates in France (Gury 2011), Italy (Di Pietro 2004), South Africa (Bokana 2010) and the UK (Johnes and McNabb 2004).

P. Jia · T. Maloney (✉)
Economics Department, Faculty of Business and Law, Auckland University of Technology, Private Bag 92006, Auckland 1142, New Zealand
e-mail: tim.maloney@aut.ac.nz

more common as less able or academically prepared students are admitted to university. Public funding authorities are also increasingly concerned by the potential waste of public expenditures on students who subsequently fail at university. For example, reducing non-completion rates are a core concern of recent reforms of the tertiary education sector in New Zealand (e.g., see New Zealand Ministry of Education 2004).

There is a substantial body of empirical literature on the determinants of university non-completion outcomes (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; Rask 2010; Belloc et al. 2010). Although a comprehensive understanding of the relative importance of the various reasons for non-completion behavior remains elusive, it has been widely recognized that individual characteristics, student educational and family backgrounds, and institutional factors are the main determinants of these outcomes. However, due mainly to limited data availability, most previous studies have utilized relatively few factors in this analysis. Using a more comprehensive dataset, our study is able to analyze the impact of a wide variety of explanatory variables on poor university outcomes.

Our paper uses administrative data from a large public university in New Zealand to estimate the determinants of course non-completion in the first year and university non-retention in the second year. Administrative data have a number of advantages for the purposes of this study. Firstly, these data are gathered as part of the normal application process and thus no additional expense or inconvenience is incurred in acquiring this information. Secondly, because these data are collected for enrollment purposes, the variables and their definitions are consistent over time. This is an important aspect if we want to use historical data to predict the at risk status of future students on an ongoing basis.

This study has two goals. Firstly, we want to estimate the effects of specific factors that may lead to both first-year course non-completions and second-year student non-retentions. This work largely parallels previous studies in this area. Secondly, and more innovatively, we use these results to test the efficacy of a potential predictive risk tool for the early identification of students who are most vulnerable to adverse outcomes at university. This is a trial to show how existing administrative data could be used to target intervention services (e.g., special tutorials or classes, student advising or mentoring services) at the most at risk students entering university for the first time. We see this as a major contribution of this study.

The rest of the paper is organized as follows. The "Literature review" section provides a brief overview of the relevant literature in this area. The "Data and methodology" section describes the data used in our regression analysis and summarizes our econometric approach. The "Empirical results" section analyses our regression results. Finally, the "Conclusion" section draws conclusions from this analysis, and suggests possible directions for future work in this area.

Literature review

Predictive risk analysis has been used previously in areas such as health care and child protection (e.g., see Billings et al. 2012; Vaithianathan et al. 2013). To our knowledge, this approach has not been formally applied to the analysis of students at risk of adverse academic outcomes at university. The key is to assess the overall predictive power of the regression analysis and its efficacy as a predictive tool with a separate 'validation' sample.

There is a substantial literature estimating the factors that influence poor student university experiences. For example, many studies have shown significant effects of student demographic characteristics (e.g., ethnicity, gender, country of origin and age) on dropout behavior (e.g., 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). Prior academic performance has been found to be related to success at university (e.g., see Betts and Morell 1999; Cohn et al. 2004; Cyrenne and Chan 2012; Ficano 2012) or the choice of fields of study (e.g., Rask 2010). Other studies have empirically examined the impact of past academic performance and other factors on student dropout behavior (e.g., Wetzel et al. 1999; Montmarquette et al. 2001; Di Pietro 2004; Johnes and McNabb 2004; Singell 2004; Bai and Maloney 2006; Ishitani 2006; Stratton et al. 2008; Belloc et al. 2010; Ost 2010; Gury 2011).

Although there is a substantial literature on the effects of class size on school academic performance (e.g., see Angrist and Lavy 1999; Krueger 2003; Rivkin et al. 2005; Fredriksson et al. 2014), to our knowledge, no previous published work has considered the effects of class size on academic outcomes at university. We use non-experimental data in our study to estimate the effects of various facets of class size on the probability of course non-completion and university non-retention.

Past research confirms the considerable differences of study areas on student dropout behavior (e.g., Robst et al. 1998; Rask 2010; Rodgers 2013). Students who study science or engineering may be more likely to drop out than those who study arts or business, possibly due to the degree of difficulty of course material and academic expectations in these programs.

Data and methodology

Administrative data were provided by a large public university in New Zealand for the purposes of this study.[2] Data were made available on all first-year students who enrolled in university Bachelor degree programs for the first time during the 2009 through 2012 academic years. The full sample contains 15,833 individuals and 88,464 course-specific observations. Individual student observations are used to examine non-retention outcomes in the second year, while individual course observations are used to investigate course non-completion outcomes in the first year.

Variable definitions and descriptive statistics are provided in Table 1. Our dataset contains detailed information typically available at the time of initial enrollment at university (e.g., year of entry, demographic characteristics, high school academic performance, and course and program enrollment information).

Two dependent variables are used in this study: course non-completion outcomes in the first year and university non-retention outcomes in the second year.[3] The first dummy variable is set equal to one if the student did not successfully complete a course (i.e., receive a passing grade) in the first year; zero otherwise. The second dummy variable is set equal to one if the student did not return to re-enroll at this university at the beginning of

[2] There are only eight universities in New Zealand. All of them are publically funded. The average enrolment level was slightly less than 17,000 in 2012. The University of Auckland was the largest with over 32,000 students and Lincoln University was the smallest with less than 4,000 students.

[3] We do not distinguish in this analysis between course dropouts (i.e., individuals who discontinued study prior to the end of the semester) and course failures (i.e., individuals who continued to the end of the semester, completed all assessments, but failed the course). This is largely because of the government reporting requirements in New Zealand that emphasize non-completion outcomes as a result of either process.

Table 1 Descriptive statistics and variable definitions from full samples ($n = 88{,}464$ course observations and $n = 15{,}833$ student observations)

Variable	Definition	Mean (SD)
Dependent variables		
Non-completion	1 if the student does not successfully complete a first-year course; zero otherwise	0.157 (0.364)
Non-retention	1 if the student does not return to university in the second year; zero otherwise	0.229 (0.420)
Year of cohort		
Year 2009	1 if student first enrolls in year 2009; zero otherwise	0.231 (0.421)
Year 2010	1 if student first enrolls in year 2010; zero otherwise	0.252 (0.434)
Year 2011	1 if student first enrolls in year 2011; zero otherwise	0.237 (0.426)
Year 2012	1 if student first enrolls in year 2012; zero otherwise	0.280 (0.449)
Ethnicity		
European	1 if student reports ethnicity as European; zero otherwise	0.421 (0.494)
Asian	1 if student reports ethnicity as Asian; zero otherwise	0.226 (0.418)
Pacifica	1 if student reports ethnicity as Pacifica; zero otherwise	0.119 (0.324)
Māori	1 if student reports ethnicity as Māori; zero otherwise	0.107 (0.310)
Other ethnicities	1 if student reports other ethnicity; zero otherwise	0.079 (0.270)
Not reported	1 if students reports no ethnicity; zero otherwise	0.047 (0.212)
Country of origin		
New Zealand	1 if student reports New Zealand as country of origin; zero otherwise	0.753 (0.431)
China	1 if student reports China as country of origin; zero otherwise	0.056 (0.229)
Korea	1 if student reports Korea as country of origin; zero otherwise	0.021 (0.144)
India	1 if student reports India as country of origin; zero otherwise	0.013 (0.115)
Other countries	1 if student reports other country of origin; zero otherwise	0.143 (0.350)
Not reported	1 if students reports no country of origin; zero otherwise	0.013 (0.115)
Personal characteristics		
Female	1 if student is female; zero if male	0.609 (0.488)
Part-time	1 if student is enrolled part-time; zero if full-time	0.273 (0.445)
Language	1 if student reports a first language; zero otherwise	0.567 (0.495)

Table 1 continued

Variable	Definition	Mean (SD)
English	1 if English is first language; zero otherwise (conditional on Language = 1)	0.769 (0.421)
Domestic	1 if student receives domestic funding; zero otherwise	0.928 (0.258)
Age	Mean age	21.958 (6.289)
High school information		
Known NCEA Score	1 if NCEA score is available from last year of school; zero otherwise	0.498 (0.500)
Actual NCEA Score	Actual NCEA score (conditional on availability of score)	158.401 (61.197)
Literacy/numeracy	1 if student took literacy and numeracy test in school; zero otherwise	0.229 (0.420)
School decile	Mean school decile (conditional on availability of school decile)	6.834 (2.812)
Entrance type		
NCEA admission	1 if student entered through NCEA results; zero otherwise	0.427 (0.495)
Special admission	1 if student entered through Special Admission category; zero otherwise	0.153 (0.360)
Internal	1 if student entered through a pre-degree program; zero otherwise	0.105 (0.307)
Cambridge/IB	1 if student entered through Cambridge or International Bachelaureate; zero otherwise	0.016 (0.127)
Other entry types	1 if student entered through some other category; zero otherwise	0.298 (0.458)
Course information		
Study hours	Recommended out-of-class course study and preparation hours	179.438 (62.350)
Known contact	1 if contact hours for the course are reported; zero otherwise	0.844 (0.363)
Contact hours	Scheduled in-class course lecture, tutorial, workshop, laboratory or meeting hours (conditional on reporting contact hours)	63.817 (40.518)
Course size	Total number of students enrolled in the course	553.058 (531.454)
Class size	Average class size in the course	38.321 (29.191)
Internet content	1 if course is supported with internet content; zero otherwise	0.583 (0.493)
Level 4	1 if course is at level 4 (pre-degree); zero otherwise	0.005 (0.065)
Level 5	1 if course is at level 5 (first year); zero otherwise	0.844 (0.363)
Level 6	1 if course is at level 6 (second year); zero otherwise	0.149 (0.356)
Level 7	1 if course is at level 7 (third year); zero otherwise	0.003 (0.058)

Table 1 continued

Variable	Definition	Mean (SD)
Individual academic information		
Number of courses	Number of courses taken by the student	5.334 (2.293)
Level 6 or 7 courses	Proportion of level 6 or 7 courses taken by the student	0.130 (0.192)
Double degree	1 if student is enrolled in a double degree; zero otherwise	0.009 (0.093)
Multiple campuses	1 if student was enrolled in courses across different campuses within the relevant period; zero otherwise	0.091 (0.287)
First-year programs of entry		
BA	1 if student enrolled in Bachelor of Arts; zero otherwise	0.083 (0.276)
BBus	1 if student enrolled in Bachelor of Business; zero otherwise	0.264 (0.441)
BCIS	1 if student enrolled in Bachelor of Computer Information Science; zero otherwise	0.044 (0.205)
BCS	1 if student enrolled in Bachelor of Communication Studies; zero otherwise	0.073 (0.260)
BDes	1 if student enrolled in Bachelor of Design	0.080 (0.271)
BEdu	1 if student enrolled in Bachelor of Education	0.042 (0.201)
BEngT	1 if student enrolled in Bachelor of Engineering Technology; zero otherwise	0.027 (0.162)
BHS	1 if student enrolled in Bachelor of Health Science; zero otherwise	0.199 (0.399)
BIHM	1 if student enrolled in Bachelor of International Hospitality Management; zero otherwise	0.042 (0.201)
BSR	1 if student enrolled in Bachelor of Sports and Recreation; zero otherwise	0.065 (0.246)
Others	1 if student enrolled in another smaller program; zero otherwise	0.089 (0.284)

the second year; zero otherwise. The results reported in Table 1 show that the mean non-completion rate is 0.157 for the course observations in our sample. The mean non-retention rate is 0.229 for the student observations in our sample. Of course, students may leave university either temporarily or permanently, and for a variety of reasons.[4]

We have data on all first-year students from four annual cohorts. Our observations are fairly evenly distributed across these 4 years (see Table 1). We have six dummy variables for a student's self-reported ethnicity (i.e., European (42.1 %), Asian (22.6 %), Pacifica (11.9 %), Māori (10.7 %), other ethnicities (7.9 %), and ethnicity not reported (4.7 %)). We have six dummy variables on country of origin. Most first-year students are from New Zealand (75.3 %), followed by China (5.6 %), Korea (2.1 %) and India (1.3 %). All other reported countries of origin are combined into a residual category accounting for 14.3 % of first-year students. Those not reporting their country of origin makeup 1.3 % of our sample.

Other personal characteristics include being female (60.9 % of our sample) and enrolling for study part-time (27.3 %). Of the 56.7 % of our first-year students who reported information on their first or primary language, 76.9 % identified English as their native language. Domestic students are defined as those receiving domestic funding status (i.e., government subsidies). They comprised 92.8 % of the first-year students at this university. The mean age of these students was 21.958.

Our dataset contains some information on the high school records of these students. Most students in New Zealand sit the National Certificate of Educational Achievement (NCEA) exams in the last 3 years of high school.[5] These are national end-of-the-year exams across a number of compulsory and optional subject areas. Our dataset includes a summary measure of the overall performance on these NCEA exams in the final year of high school. As indicated in Table 1, this NCEA score was available for about half of the first-year students ('Known NCEA Score'). NCEA results are not strictly required for university entrance in New Zealand. Older first-time domestic students may not have taken NCEA exams, which were only fully implemented in 2004. Foreign students do have to take these NCEA exams. This is because the NCEA system is not an external entry exam like the SAT or ACT exams used in the USA that would normally be taken by overseas applicants.

Two additional variables are available on the educational background of our students. A value of one for the variable 'Literacy/numeracy' indicates that tests was taken during high school to investigate possible issues over appropriate literacy and numeracy levels. In New Zealand, high schools are sorted into deciles based on the socioeconomic status of residents in the school catchment area.[6] For example, a decile 1 high school is among the 10 % of schools from poorest socioeconomic areas, while a decile 10 high school is from the wealthiest socioeconomic areas. The mean school decile in our sample is 6.834, indicating that these first-year students were drawn predominantly from higher decile schools.

[4] Possible explanations for dropout behaviour include students struggling academically at university, transferring to other institutions, leaving for employment opportunities, etc. It should be noted that we have no information in the database on the reasons why individuals may have failed to return to this university at the beginning of the second year.

[5] For more information on the NCEA system, see http://www.nzqa.govt.nz/qualifications-standards/qualifications/ncea/.

[6] For more information on the process used to determine school deciles see http://www.minedu.govt.nz/NZEducation/EducationPolicies/Schools/SchoolOperations/Resourcing/OperationalFunding/Deciles/HowTheDecileIsCalculated.aspx.

Acquiring University Entrance status is advertised as the minimum requirement for gaining entry to Bachelor degree programs in New Zealand universities. This normally consists of receiving sufficient credits in NCEA exams in the final year of high school in at least three approved subject areas. Less than half of the first-year students in our sample (42.7 %) gained admission to this university through their NCEA scores. A small minority of students (1.6 %) gained University Entrance through the completion of the more prestigious and challenging Cambridge or International Baccalaureate programs at secondary school ('Cambridge/IB'). Yet, there are three additional ways in which first-year students could gain entry to this university.[7] The category 'Other Entry Types' primarily includes foreign students who obtained admission through equivalent overseas high school qualifications. They comprised 29.8 % of all first-year students in our sample. Students can enter New Zealand universities through a 'Special Admission' status which refers to those who did not receive University Entrance but were admitted because of their age and relevant experiences.[8] Special Admissions accounted for 15.3 % of our sample. Finally, 'Internal' status refers to the 10.5 % of students who were granted enrollment because they had successfully completed a 'pre-degree' certificate or diploma at this university. They are a particularly interesting group for this analysis. These are students without University Entrance who needed to complete a 'bridging program' that consists of basic literacy, numeracy, and other academic subjects to prepare them for entry into the higher level Bachelor degree programs.[9]

For the purpose of analyzing course non-completion outcomes, our administrative data contain some potentially useful information on the characteristics of these courses. We know the recommended 'Study Hours' from recorded course outlines. These are the expected number of hours for study and assessment preparation outside of regular class time. Most courses (84.4 %) report 'Known Contact' hours.[10] These contact hours include scheduled lecture, tutorial, workshop, and lab hours (i.e., regular in-class time). They could also include scheduled office hours, and group study hours and generic academic preparation workshops in areas such as English, writing skills, and mathematics.

We also know the 'Course Size' and average 'Class Size'. The former is the total number of students enrolled in the course, where the latter is the average number of students in a classroom. For example, a large first-year course could have 1,000 students enrolled. This would be the course size. These students could be taught in a single large class of 1,000, or they could be taught in 20 classes with an average class size of 50 students. We consider the separate effects of both course and class size on course non-completion outcomes. We also know whether or not the course is supported with internet

[7] We excluded all students from our sample who had studied previously at another university. These 'university transfers' may not be directly comparable to students who enter university Bachelor degree programmes for the first time.

[8] Universities consider a range of factors in granting this Special Admission status. The applicant must be at least 20 years old by the time that they enroll at university. Previous educational records, work and training experiences, English language skills, and motivations for study are also considered.

[9] Evidence from our sample suggests that those entering university through this Special Admission status are at higher risk of poor university outcomes on a number of dimensions compared to those with University Entrance. They are relatively more likely to be Pacifica or Māori (27.9 vs. 20.0 %), male (41.4 vs. 34.8 %), studying part-time (45.8 vs. 17.9 %), and originally from schools in the bottom three deciles (13.7 vs. 11.9 %).

[10] It is unclear why contact hours were not reported for some of these courses. This could be related to the nature of some of these courses. For example, a course may have less formal scheduled contact hours if it involves largely 'independent study'.

content. Finally, we know the academic level of the course. A 'Level 4' course contains content that is intended for students below a Bachelor degree level. Most courses taken by these first-year students (84.4 %) are intended for the first year of university study ('Level 5'), but some students enroll in courses intended for second- and third-year study ['Level 6' (14.9 %) and 'Level 7' (0.3 %), respectively].[11]

We have additional academic information on students including the number of courses in which they had enrolled ('Courses Taken') and the proportion taken at Levels 6 or 7 ('Level 6/7 Courses'). We also know whether or not the student had enrolled for a double-degree program, and whether or not courses had been held across multiple campuses at this university. Less than 1 % of first-year students enrolled for a double degree, partly due to stringency of the entry requirements. It was also possible for classes at this university to be held across as many as three distinct campuses. Because the movement across the campuses in different geographic locations within the metropolitan area could have resulted in some disruption to the learning process, we wanted to test whether or not this factor had any impact on the probabilities of course non-completion and student non-retention.

Finally, our dataset contains information on the initial program of study. We use dummy variables to identify the largest 10 Bachelor degree programs. The residual category includes all of the smaller degree programs (8.9 % of students in our sample). The largest three programs are the Bachelor of Business (26.4 %), the Bachelor of Health Sciences (19.9 %), and the Bachelor of Arts (8.3 %).

Maximum likelihood probit analysis will be used to estimate the effects of these various factors on our two dummy dependent variables. The basic probit model can be written:

$$Y_i^* = \boldsymbol{\beta X}_i + u_i \tag{1}$$

where Y_i^* is a latent variable associated with course non-completion or student non-retention propensities. What we observe is a dummy variable Y_i that equals 1 if the course was not successfully completed in the first year, or the student did not return to re-enroll at this university in the second year; zero otherwise. This depends on the latent dependent variable crossing an arbitrary threshold of zero.

$$Y_i = \begin{cases} 1 & \text{if} \quad Y_i^* > 0 \\ 0 & \text{if} \quad Y_i^* \leq 0 \end{cases} \tag{2}$$

All of the aforementioned factors are included in the vector $\boldsymbol{X_i}$. The unknown coefficients are represented by the $\boldsymbol{\beta}$ vector which will need to be estimated. The probability of course non-completion or student non-retention can be denoted as:

$$P(Y_i^* > 0) = P(\boldsymbol{\beta X}_i + u_i > 0) = P(u_i > -\boldsymbol{\beta X}_i) = \boldsymbol{\Phi}(\boldsymbol{\beta X}_i) \tag{3}$$

where $\boldsymbol{\Phi}(\cdot)$ is the cumulative distribution function (CDF) of the standard normal.

We will use the average marginal effects to describe the influence of a one-unit change in an explanatory variable on these probabilities. This is because a probit model is a nonlinear function of the coefficients, and the marginal effects are dependent on the values of the other regressors. For any particular factor X_k, this partial derivative can be written:

[11] The typical university baccalaureate programme in New Zealand is completed in 3 years of full-time study.

$$\frac{\partial P(Y_i = 1|\boldsymbol{X}_i)}{\partial X_k} = \beta_k \phi(\boldsymbol{\beta X}_i) \tag{4}$$

where $\phi(\cdot)$ is the probability distribution function (PDF) of the standard normal.[12]

This probit estimation is also used in the development of our predictive risk models (PRMs), which then can be used to generate risk scores for any first-year student enrolling at this university. To assess the effectiveness of these predictive risk tools, we compare our predicted outcomes against the actual outcomes for each of our dependent variables. For this reason, our full sample is randomly split into two equal-sized 'estimation' and 'validation' samples (e.g., see medical applications of this methodology in Billings et al. 2012). The estimation samples will be used to estimate the probit models, and the validation samples will be used to assess how well the PRMs correctly identify the actual course non-completion outcomes and student non-retention outcomes.

Empirical results

Two separate maximum likelihood probit models were estimated for this study, using the estimation samples for course non-completion outcomes in the first year and student non-retention outcomes in the second year. The estimated coefficients, standard errors, and average marginal effects are presented in Table 2 for both dependent variables.

Results on first-year course non-completion

Holding constant other measured factors, we find that the probability of course non-completion varies systematically across the years (where 2012 is the excluded category or benchmark year). All three estimated coefficients on the included year dummies are negative and statistically different from zero at better than a 5 % level. This says that the course non-completion probability was highest in 2012 compared to the previous 3 years. However, 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 course non-completion rates over time.

Ethnicity appears to have a substantial impact on the probability of course non-completion. Pacifica and Māori students have significantly higher probabilities of course non-completions. The estimated partial derivatives indicate their probabilities of not successfully completing a course are 10.730 and 6.971 percentage points higher than those for the omitted category of European students, respectively. Students with other ethnicities and ethnicities not reported are also relatively more likely to have experienced a course non-completion. Yet, there is no statistical evidence of a difference in course non-completion rates between observationally equivalent Asian and European students.

The estimated coefficients on the first four dummy variables on country of origin are positive, but generally not statistically different from zero. The only exception is that students from Korea have a course non-completion rate that is 2.819 percentage points higher than that of the omitted category of students from New Zealand. This estimated effect is statistically significant at a 10 % level. However, we find that students who did not

[12] The marginal effects could also be calculated at the sample means for the explanatory variables. For continuous functions in large samples, this technique yields similar results to the sample mean for the individual marginal effects.

Table 2 Estimated results from maximum likelihood probit analysis on course non-completions in the first year and student non-retentions in the second year for the estimation samples

Variable	Probability of course non-completion			Probability of student non-retention		
	Coefficient	Std. error	*dy/dx*	Coefficient	Std. error	*dy/dx*
Constant	−0.8337***	0.1256	–	−0.5664***		0.1906
Year of cohort						
Year 2009	−0.1121***	0.0334	−2.308***	−0.0574	0.0485	−1.531
Year 2010	−0.0716**	0.0305	−1.473**	0.1520***	0.0454	4.057***
Year 2011	−0.2000***	0.0308	−4.118***	0.0361	0.0464	0.964
Ethnicity						
Asian	0.0506	0.0357	1.043	−0.2082***	0.0523	−5.556***
Pacifica	0.5211***	0.0408	10.730***	0.0185	0.0591	4.927
Māori	0.3385***	0.0382	6.971***	0.2136***	0.0558	5.700***
Other ethnicities	0.1212**	0.0485	2.496**	−0.1309	0.0698	−3.494
Not reported	0.1915**	0.0865	3.942**	−0.1737	0.1273	−4.635
Country of origin						
China	0.0647	0.0682	1.332	−0.0918	0.0991	−2.449
Korea	0.1369*	0.0819	2.819*	0.1780	0.1264	−4.750
India	0.0450	0.1049	0.927	−0.0688	0.1557	−1.837
Other countries	0.0152	0.0391	0.312	−0.1187**	0.0563	−3.169**
Not reported	0.5176***	0.0943	10.659***	0.4594***	0.1349	12.259***
Personal characteristics						
Female	−0.1332***	0.0260	−2.743***	−0.0402	0.0375	−1.072
Part-time	0.8368***	0.0281	17.231***	0.6758***	0.0505	18.037***
Language	0.0471	0.0405	0.970	0.0593	0.0601	1.584
English	−0.0516	0.0436	−1.063	−0.0520	0.0641	−1.387
Domestic	0.2162***	0.0779	4.452***	0.1140	0.1134	3.041
Under age 18	−0.1546	0.1518	−3.183	−0.1284	0.2073	−3.428

Table 2 continued

Variable	Probability of course non-completion			Probability of student non-retention		
	Coefficient	Std. error	*dy*/*dx*	Coefficient	Std. error	*dy*/*dx*
Age 19	0.0011	0.0326	0.022	0.0451	0.0491	1.203
Age 20	0.1146***	0.0399	2.359***	0.0116	0.0604	0.308
Age 21	0.1386***	0.0484	2.854***	0.0417	0.0708	1.111
Age 22	0.0504	0.0565	1.037	−0.0243	0.0818	−0.649
Age 23	0.0834	0.0625	1.718	0.0380	0.0903	1.014
Age 24	−0.1046	0.0752	−2.154	−0.1584	0.1096	−4.226
Age 25	−0.0353	0.0842	−0.727	0.1248	0.1209	3.331
Ages 26–30	−0.1339**	0.0572	−2.757**	0.0383	0.0788	1.021
Ages 31–35	−0.1879**	0.0835	−3.870**	−0.0341	0.1050	−0.911
Ages 36–45	−0.2389***	0.0760	−4.919***	−0.0003	0.1036	−0.009
Ages 46+	−0.2781**	0.1171	−5.726**	0.0972	0.1410	2.593
High school information						
Known NCEA Score	0.3393***	0.0554	6.987***	0.1284	0.0819	3.425
Actual NCEA Score	−0.0045***	0.0004	−0.093***	−0.0020***	0.0005	−0.053***
Literacy/numeracy	0.1070***	0.0302	2.204***	0.1892***	0.0521	5.050***
School decile 1	0.3373***	0.0740	6.945***	0.2415**	0.1091	6.445**
School decile 2	0.1780**	0.0724	3.665**	0.0858	0.1065	2.229
School decile 3	0.1506**	0.0634	3.102**	0.0887	0.0963	2.367
School decile 4	0.0910	0.0578	1.874	0.0348	0.0865	0.930
School decile 5	0.1070	0.0675	2.202	0.0629	0.0969	1.678
School decile 7	0.0897	0.0586	1.846	0.1069	0.0870	2.852
School decile 8	0.0942	0.0603	1.939	0.1409	0.0878	3.761
School decile 9	0.1092**	0.0549	2.249**	0.0902	0.0810	2.407
School decile 10	0.1134**	0.0522	2.335**	0.0677	0.0772	1.807
Decile not reported	−0.0379	0.0605	−0.779	0.1041	0.0858	2.779

Table 2 continued

Variable	Probability of course non-completion			Probability of student non-retention		
	Coefficient	Std. error	*dy/dx*	Coefficient	Std. error	*dy/dx*
Entrance type						
Special admission	−0.0928*	0.0542	−1.912*	0.0657	0.0787	1.753
Internal	−0.1603***	0.0541	−3.301***	−0.1134	0.0789	−3.027
Cambridge/IB	−0.4781***	0.1038	−9.844***	−0.2030	0.1593	−5.417
Other entry types	−0.1869***	0.0555	−3.848***	0.0683	0.0802	1.823
Course information						
Study hours	−0.0009	0.0075	−0.018	–	–	–
Known contact	−0.1098***	0.0395	−2.262***	–	–	–
Contact hours	−0.0001	0.0005	−0.002	–	–	–
Course size/100	−0.0133***	0.0030	−0.273***	–	–	–
Class size/100	0.0288	0.0326	0.593	–	–	–
Internet content	0.0182	0.0203	0.374	–	–	–
Level 4	−0.1952*	0.1172	−4.018*	–	–	–
Level 6	−0.0048	0.0266	−0.099	–	–	–
Level 7	−0.2659*	0.1446	−5.476*	–	–	–
Individual academic information						
Number of courses	–	–	–	−0.0063	0.0124	−0.168
Level 6 or 7 courses	–	–	–	−0.8452***	0.1083	−22.555***
Double degree	−0.1092	0.1287	−2.248	−0.0366	0.2300	−0.976
Multiple campuses	−0.0148	0.0388	−0.304	0.0114	0.0624	0.305
First-year entry programs						
BBus	−0.2685***	0.0616	−5.528***	−0.5650***	0.0806	−15.078***
BCIS	−0.0253	0.0596	−0.522	−0.3876***	0.0990	−10.344***
BCS	−0.3385***	0.0603	−6.970***	−0.6136***	0.0954	−16.376***
BDes	−0.4325***	0.0573	−8.906***	−0.4338***	0.0887	−11.577***

Table 2 continued

Variable	Probability of course non-completion			Probability of student non-retention		
	Coefficient	Std. error	*dy*/*dx*	Coefficient	Std. error	*dy*/*dx*
BEdu	−0.5670***	0.0741	−11.676***	−0.7268***	0.1082	−19.398***
BEngT	−0.0534	0.0741	−1.100	−0.4847***	0.1156	−12.935***
BHS	−0.4096***	0.0485	−8.434***	−0.3812***	0.0668	−10.174***
BIHM	−0.3142***	0.0639	−6.470***	−0.3809***	0.0977	−10.165***
BSR	−0.2050***	0.0516	−4.222***	−0.2937***	0.0846	−7.839***
Other progs	−0.0257	0.0490	−0.529	−0.4115***	0.0801	−10.981***
Pseudo R^2	0.1429				0.1146	
Pseudo log-likelihood	−16,543.7				−3,771.6	
Area under the ROC curve	0.7579				0.7179	
N	44,232				7,917	

The 50 % estimation samples used in these regressions were randomly selected from the full samples on course non-completions and student non-retentions. The estimated standard errors are adjusted for the clustering of course observations for individual students in the first regression. The columns labelled ‘*dy*/*dx*’ contain the estimated mean marginal effects (i.e., the estimated percentage point changes in the probabilities of course non-completions or student non-retentions for one-unit changes in the independent variables)

*** Significance at the 1 % level

** Significance at the 5 % level

* Significance at the 10 % level

report their country of origin have course non-completion rates that are 10.659 percentage points higher than those for otherwise observationally equivalent New Zealand students. This estimated effect is statistically significant at better than a 1 % level. In other words, those not reporting a country of origin at the time of initial university enrollment appear to be the highest risk group within this category of variables for not successfully completing their first-year courses.

Consistent with earlier studies, female students have a relatively lower estimated probability of course non-completion. Holding other things constant, being female lowers this probability of course non-completion by 2.743 percentage points. This effect is statistically significant at better than 1 % level.

Studying part-time is estimated to substantially increase the rate of course non-completion. Being a part-time student increases this probability by 17.231 percentage points. English as the first language has no measurable impact on course non-completion rates in our sample. Being a domestic student (i.e., receiving government subsidies) increases the probability of course non-completion by 4.452 percentage points.

We use a series of dummy variables to allow for flexibility in the age effects on course non-completions. One dummy variable is used for being under the age of 18. A series of dummies are used for individual ages from 18 to 25 inclusive, and four dummies are used for age ranges 26 through 30, 31 through 35, 36 through 45 and 46 years and older. The omitted age group is the modal group of 18 year olds. Ages of 20 and 21 have positive and significant effects on the probability of course non-completion. Their probabilities of course non-completion are estimated to be, respectively, 2.359 and 2.854 percentage points higher than those of the typical student entering this university at age 18. Ages above 25 have negative and statistically significant effects on this probability of course non-completion. Although older students are generally less at risk of not successfully completing their first-year courses, those in the relevant age range for Special Admission entry status may be particularly vulnerable.

Students who scored higher on their NCEA exams are found to be at lower risk of course non-completion during their first year at university. Two variables must be considered in interpreting these results. The first is a dummy variable on having information on these exam results (Known NCEA Score), and the second is the composite exam score from this last year of high school (Actual NCEA Score). The estimated effect of having an NCEA score on this probability of course completion could be written:

$$\frac{\partial P(\mathrm{Non-Completion})}{\partial \mathrm{Known\ NCEA\ Score}} = 6.987 - 0.093 \times \mathrm{Actual\ NCEA\ Score} \tag{5}$$

We know from Table 1 that the sample mean for those reporting a NCEA score is 158.401. Thus, for the average student with NCEA results, these exams reduce the probability of course non-completion in the first year by an average of 7.744 percentage points:

$$\frac{\partial P(\mathrm{Non-Completion})}{\partial \mathrm{Known\ NCEA\ \ Score}} = 6.987 - 0.093 \times 158.401 \approx -7.744 \tag{6}$$

The previous section indicated that the dummy variable on literacy/numeracy tests in high school picks up possible concerns over the reading, writing, and mathematics skills for students. As expected, taking these tests is associated with a significant increase in 2.204 percentage points in the average probability of course non-completion.

We expected that students from lower school deciles would have higher probabilities of paper non-completions during their first year at university. This result is largely confirmed by our analysis, but some discussion around these findings is needed. Firstly, the omitted category includes students from decile 6 schools. Schools in the bottom three deciles all have positive and statistically significant effects on the probability of course non-completion relative to this benchmark group. These partial derivatives rise in magnitude as we move to schools from increasingly poorer socioeconomic areas (3.102 percentage points for decile 3, 3.665 percentage points for decile 2 and 6.945 percentage points for decile 1). Yet, there is no statistical evidence of improved outcomes as we move to schools from better socioeconomic areas. In fact, schools in the top two deciles also have positive and statistically significant effects on the probability of course non-completion.

The regression results from the dummy variables on school deciles suggest a possible 'U shaped' relationship between these deciles and course non-completion rates. To formally test for this possibility, we substituted a continuous measure of school deciles (integers from 1 to 10) and its squared value in place of these dummy variables in an auxiliary regression. If this U-shaped relationship exists, we should find a negative coefficient on this continuous school decile measure and a positive value on its squared value. The full regression results are not reported but are available from the authors on request. However, the mean marginal effects had the expected signs and were individually statistically significant at better than a 1 % level. To interpret these results, we can take the partial derivative of this probability of course non-completion with respect to the school decile:

$$\frac{\partial P(\mathrm{Non-Completion})}{\partial \mathrm{School\ Decile}} = -1.923 + 0.143 \times 2 \times \mathrm{School\ \ Decile} \tag{7}$$

If we set the partial derivative above equal to zero and solve for this school decile, we get a value of approximately 6.724. This says that observationally equivalent students from school deciles either below or above this value have higher probabilities of not successfully completing their first-year courses (although the students from lowest decile schools are still relatively more at risk than those from the highest decile schools). One possible explanation for these results is that many students at this university who came from the highest decile schools were unable to gain admittance to higher ranked universities in New Zealand and overseas and generally did not have the same academic preparation (or motivation) of students from mid-decile schools.

All of the included university entrance types have negative and statistically significant effects on the probability of course non-completion in the first year. The omitted category is the conventional NCEA Admission. Since this entrance type is closely connected to the effects of the NCEA Score discussed earlier, this admission category is not likely to be a risk indicator because NCEA scores for these students reduce the probability of course non-completion. Yet, we can say that students entering university with a Cambridge or International Baccalaureate qualification are at the lowest relative risk. This entrance type reduces the probability of course non-completion by an average of 9.844 percentage points compared to this reference group. Both Internal and Special Admission entry are of particular interest in this study, because these are students who do not have the academic backgrounds to gain direct entry to this university. Internal entry is granted to students who successfully complete a pre-degree qualification at this university, while Special Admission entry is given to students who reach the age of 21. We find that the Special Admission group is relatively more at risk of course non-completions compared to younger students who successfully completed this bridging program. This estimated effect on Special Admission entry combined with at risk nature of students aged in their early twenties makes this a particularly vulnerable group.

The remaining covariates in this regression model relate to the courses or degree programs in which these students were enrolled during their first year at university. As mentioned previously, we draw a distinction between the overall number of students enrolled in a course (course size) and the average number of students in a classroom (class size). To ease the interpretation of the estimated results, both variables are divided by 100. Individuals often enroll in large first-year courses, but these can be taught in either large settings (e.g., a single mass lecture) or small settings (e.g., multiple streams taught in smaller classrooms). These course and class size effects could be quite different for the probability of course non-completion. For example, courses with large enrollments could reduce the probability of course non-completion because of the introductory nature of the subject material and the need for large-scale assessments. On the other hand, and similar to the usual justification in the literature on class size effects in schools, large classroom settings could increase course non-completions due to the lack of individual attention for students. These are precisely the direction of the effects that we find in our analysis. The estimated course size effect is negative and statistically significant at better than a 1 % level. We find that an increase in course enrollment of 100 students would reduce the probability of course non-completion by 0.273 percentage points. The estimated class size effect is positive, but not statistically significant at conventional test levels. A positive sign on class size would suggest that course non-completion rates might be reduced by enrolling students in large first-year courses, but teaching them in smaller classroom settings.

Finally, our results indicate that program study areas play an important role in course non-completion outcomes. We know the degree programs in which these students initially enrolled, which include multiple programs for those doing a double degree. Relative to the reference group of students in Bachelor of Arts program, all of the other degree programs had negative effects on the probability of course non-completions in the first year. The three with the lowest course non-completion rates were in the Bachelor of Education (BEdu −11.676 percentage points), the Bachelor of Design (BDes −8.906 percentage points), and the Bachelor of Health Sciences (BHS −8.434 percentage points). The three programs with the highest course non-completion rates (other than the omitted Bachelor of Arts) were the Bachelor of Computer Information Science (BCIS −0.522 percentage points), the other smaller degree programs (Other Progs including Mathematics −0.529), and the Bachelor of Engineering Technology (BEngT −1.100 percentage points). All three of these programs had course non-completion rates that were insignificantly different from the Bachelor of Arts reference group. Some caution should be exercised in interpreting these results. They could indicate something about the rigor or difficulty of first-year study in these areas, but could equally indicate something about the unobserved characteristics of the students who enroll in these degree programs. However, these results are largely consistent with the findings reported by Rask (2010) that grades and student retention rates are systematically lower in the STEM subjects of Sciences, Computing, Mathematics and Engineering.

Results on second-year university non-retention

The last three columns of Table 2 report the regression results on student non-retention outcomes in the second year for our estimation sample. Recall that both Pacifica and Māori students were significantly more likely to not complete their first-year courses. Compared to the omitted ethnic group of Europeans, the only ethnic group with a statistically significant positive effect on student non-retentions is Māori. Specifically, Māori students have a non-retention probability that is, on average, 5.700 percentage points higher than that of European students. This suggests that Pacifica students are the most likely ethnic group to not complete their courses in the first year, while Māori students are the most likely ethnic group to not return to the university in the second year. Only Asian students have a statistically significant negative effect. Relative to Europeans, Asian ethnicity reduces the probability of student non-retention by 5.556 percentage points.

The estimated results for those not reporting a country of origin have a similar positive and significant effects for both student non-retention and course non-completion. This suggests that students not reporting a country of origin are in the highest risk group for both adverse outcomes. Female students are at lower risk of both course non-completion and non-retention. However, the latter effect is not statistically significant. Part-time students are substantially more likely to drop out of university in the second year. Studying part-time increases the probability of non-retention by an average of 18.037 percentage points. Thus, part-time study is arguably the single most important single at risk factor for poor university outcomes. There is no statistical evidence of any impacts of age on student retention, in contrast to measurable effects of age on course non-completion.

Students who had higher NCEA exam scores were less likely to drop out of university in the second year. Again, we need to estimate this overall impact using the two estimated average marginal effects:

$$\frac{\partial P(\text{Non} - \text{Retention})}{\partial \text{Known NCEA Score}} = 3.425 - 0.053 \times \text{Actual NCEA Score} \tag{8}$$

We estimate that the probability of student non-retention declines by 4.970 percentage points for those with the average conditional NCEA score in the sample.

$$\frac{\partial P(\text{Non} - \text{Retention})}{\partial \text{Known NCEA Score}} = 3.425 - 0.053 \times 158.401 \approx -4.970 \tag{9}$$

We find that literacy/numeracy tests in high school are associated with a significant increase in 5.050 percentage points in the probability of non-retention in the second year at university. Recall that there were measurable impacts on the course non-completion rates at both the lower and upper school deciles. Only school decile 1 has a statistically significant effect on student non-retention rates. Being from a school in the bottom decile is associated with a 6.445 percentage point increase in the probability in non-retention in the second year, relative to the omitted category of decile 6 schools. None of the coefficients on the other school deciles are statistically different from zero. Thus, students from schools in the lowest decile have both the highest rates of course non-completion in the first year and non-retention in the second year. They are a particularly at risk group.

Students who enroll in a larger proportion of Level 6 or 7 courses have significantly lower non-retention rates in the second year. This could be because either these students are admitted to these more advanced courses because of higher unobserved academic abilities or because their enrollment in these more advanced courses signals a greater academic commitment on their part.

Finally, because the omitted group enrolling in the Bachelor of Arts program had the highest student non-retention rates, all of the estimated effects on the other programs are negative and statistically significant at better than a 1 % level. The remaining relative relationships between course non-completion and student non-retention rates in these results are not obvious. However, the Bachelor of Education, which had the lowest course non-completion rate in the first year, also had the lowest student non-retention rate in the second year.

Assessing the predictive power of our PRMs

One way to assess the overall performance of our probit regression models is to consider the Pseudo R^2 statistics reported at the bottom of Table 2. The usual interpretation is that our models can explain approximately 14.29 % of the variation in course non-completion outcomes in the first year and 11.46 % of the variation in student non-retention outcomes in the second year. These statistics, of course, only summarize the predictive power of our analysis *within* these estimation samples. We want to know how well these models perform in predicting these outcomes *outside* of these samples.

We report the area under the receiver operator characteristic (ROC) curves for both course non-completion and student non-retention outcomes in the summary statistics of Table 2 using these respective validation samples. The ROC curves characterize the relationship between the 'sensitivity' and 'specificity' in these two models. Sensitivity is the probability that a course failure (or student dropout) outcome is correctly identified. Specificity is the probability that a course completion (or student retention) outcome is correctly identified. We graphically illustrate the trade-offs between sensitivity and one minus the specificity at all possible thresholds. These results are shown in Figs. 1 and 2.

The area under the ROC curve for course non-completion is 0.7579. This indicates that there is a 75.79 % probability that a randomly selected course observation with a non-completion outcome will receive a higher risk score from our predictive risk model (PRM) than a randomly selected course observation with a completion outcome. This is an indicator of the 'target effectiveness' of this predictive risk tool could be compared to the results from other types of analyses. Similar interpretations can be given for the non-retention analysis with the area under ROC curve at 0.7179.

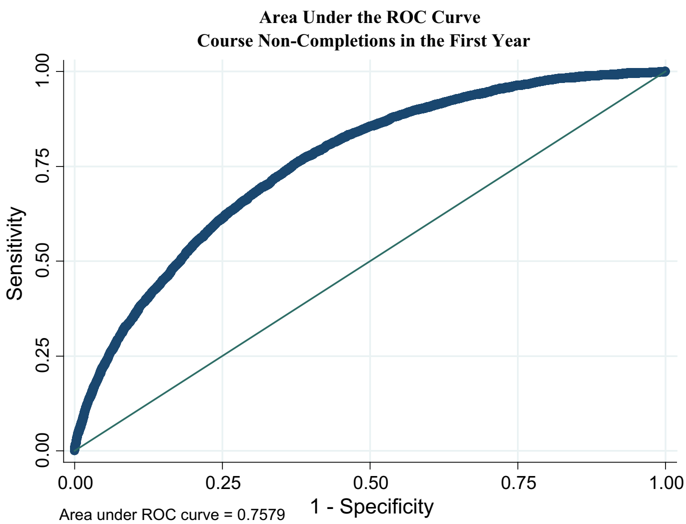

Fig. 1 Area under the ROC curve course non-completions in the first year

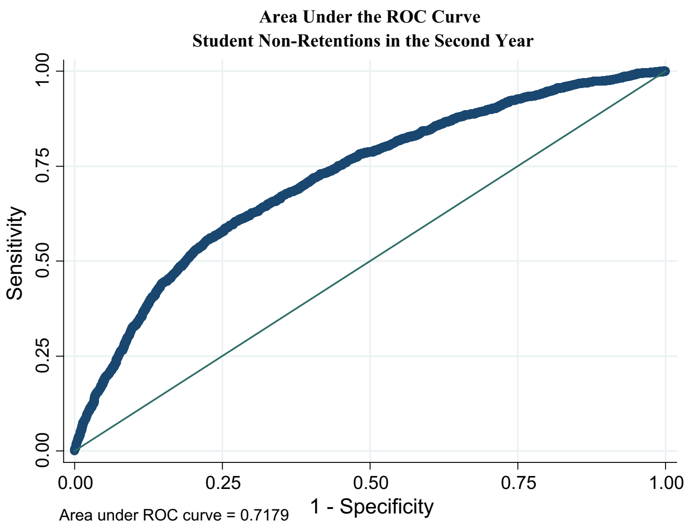

Fig. 2 Area under the ROC curve student non-retentions in the second year

Another approach is assessing the effectiveness of our PRMs is to compare predicted to actual outcomes. We use the regression results reported in Table 2 to compute risk scores for all course non-completion and student non-retention outcomes in our validation samples. We then rank these predicted probabilities, sort them into deciles, and determine the proportion of actual adverse outcomes that would be captured at every decile. Suppose, we wanted to intervene (i.e., provide specific services) to students in the top decile (i.e., those with the highest 10 % of risk scores). If our models were completely ineffective at predicting these outcomes, then the top 10 % of risk scores would account for only 10 % of the actual adverse outcomes. The results in Table 3 indicate that the highest 10 % of risk scores in the validation samples would capture 27.67 % of actual course non-completions and 21.97 % of actual student non-retentions. If we targeted the top two deciles, we would capture 45.45 % of course non-completions and 41.35 % of student non-retentions.

It is often difficult to provide any meaningful relative comparisons to the predictive power analysis of a particular PRM. Fortunately, in this situation, we had information on an existing risk analysis tool developed by this university, which provides a convenient benchmark. The university had previously used the results from a survey administered to first-year students to predict who would likely experience academic difficulties over the first year of study. University administrators attached 'weights' to the survey responses

Table 3 Percentage of outcomes correctly identified in the validation samples

	Course non-completion in first year	Student non-retention in second year
Top 1 decile (top 10 %)	27.67 %	21.97 %
Top 2 deciles (top 20 %)	45.45 %	41.35 %
Area under the ROC curve	0.7579	0.7179
N	44,232	7,916

based on subjective assessments on the relative importance of these various factors and not on a formal statistical analysis of the relationships between these variables and course non-completion outcomes.

We constructed the risk scores from our validation sample using this existing administrative tool and compared these predicted outcomes to observed course non-completions. By any measure, the predictive power of this administrative tool was substantially inferior to our PRM. Because of 'ties' in adding up these risk measures using the integer weights, we can not select only the highest 10 % of risk scores. The approximate 'top decile' using the university's administrative tool accounted for 11.78 % of course outcomes, and these captured 23.51 % of actual course non-completions in our validation sample. The top decile of risk scores using our PRM was nearly three times more likely to capture a course non-completion than the overall sample (27.67/10.00). The top decile of risk scores using the administrative tool was less than two times more likely to experience a course non-completion (23.51/11.78). In this sense, our PRM was approximately 38.63 % more 'target effective' than the existing administrative tool.

The same comparisons can be made for the top two deciles. Again, because of ties, the existing administrative tool accounted for 25.27 % of course outcomes, but captured only 39.11 % of actual course non-completions. The top two deciles of risk scores using our PRM was approximately 2.3-times more likely to experience a course non-completion than the overall sample (45.45/20.00). The top two deciles of risk scores using the administrative tool were 1.5-times more likely to experience a course non-completion (39.11/25.27). In this sense, the 'hit rate' of our PRM is approximately 46.83 % higher than the existing administrative tool.

This relatively better performance of our PRM is not that surprising given that the administrative tool used by the university had never been appropriately validated. This PRM approach has an important additional advantage. The survey-based administrative tool requires the dissemination and processing of a first-year student survey each year. This can be an expensive operation. Our PRM tool is based entirely on routine data collected as part of the enrollment process. Thus, once developed, there is virtually no additional on-going cost in using this PRM approach. In this sense, it is relatively more 'target effective' and 'cost efficient'.

Conclusion

This study has empirically estimated the determinants of course non-completion outcomes in the first year and student non-retention outcomes in the second year using administrative data from a large public university in New Zealand. These Predictive Risk Models (PRMs) have been developed to improve our understanding of the factors that that place students at risk of adverse outcomes early in their university careers. In addition, these PRMs could be used by universities to develop effective, low-cost tools for identifying students at risk of adverse outcomes and to provide early interventions for students that are most likely struggle at university.

The two dependent variables used in our regression analysis were course non-completion outcomes in the first year and student non-retention outcomes in the second year. Administrative data were taken from four annual cohorts of students entering university degree programs for the first time. Our findings suggest that a wide array of factors influence course non-completion and student non-retention probabilities. For example, part-time study is estimated to substantially raise the probabilities of both detrimental outcomes. Pacifica students are the ethnic group most at risk of course non-completion outcomes, while Māori students are the ethnic group most at risk of non-retention outcomes. Females are at lower risk of course non-completions, but not necessarily non-retentions.

Better results on national high school exams substantially reduce the risk of both course non-completions in the first and student non-retentions in the second year. We find some evidence that pre-degree, bridging programs that serve as entry points to university degree programs may be effective at reducing these adverse outcomes.

Students from high schools in the poorest socioeconomic areas have the highest course non-completion and non-retention rates. However, coming from schools in increasingly better socioeconomic areas does not necessarily improve these outcomes. In fact, students from the highest decile schools may have relatively higher course non-completion rates than those from middle decile schools. Larger overall course enrollments are associated better course outcomes. Finally, early university experiences vary substantially across the degree programs.

The areas under ROC curves were 0.7579 and 0.7179, respectively, for the course non-completion and student non-retention outcomes. The top risk decile of course observations can account for 27.67 % of actual course non-completion outcomes. The top risk decile of student observations can account for 21.97 % of actual student non-retention outcomes. These results were superior to the existing administrative tool used by this university. Our PRM is at least 38.63 % more target effective in identifying students vulnerable for course non-completions. We also claim that out PRM would also be relatively more cost-effective because it would be based on existing administrative data already collected as part of the enrollment process.

There is more that can be done in this area to better understand the determinants of these early adverse outcomes at university and to improve the accuracy of any PRM for identifying at risk students. We could improve our measures of these early university outcomes. For example, we have concentrated on the non-completion outcomes for courses. This does not distinguish between students who discontinue their study early in the semester (i.e., course dropouts) and those who do not meet the passing standards at the end of the semester (i.e., course failures). More could be done to expand the range of covariates used in the regression analysis. For example, we have no information in our administrative data on parental education, family finances, student scholarships or other financial aid, and peer and community characteristics. We could also do more with existing administrative data to improve the quality of our predictive variables. For example, we have access to only partial information on student academic performance in high school. It would be possible with available data from the Ministry of Education to gain access to the results from national exams for students over their two previous years at high school. This could greatly improve the quality of our predictive risk tool and again help the university in targeting its limited resources at the most vulnerable students.

Acknowledgments Access to the data used in this study was provided by a public university in New Zealand for the agreed purposes of this research project. The interpretations of the results presented in this study are those of the authors and do not reflect the views of this anonymous university. We thank Gail Pacheco and the anonymous referees of this journal for helpful suggestions on earlier drafts of this paper.

References

Angrist, J., & 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., & Maloney, T. (2006). Ethnicity and academic success at university. *New Zealand Economic Papers, 40*(2), 181–213.

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

Betts, J. R., & 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., & 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, 2*(4), e001667.

Bokana, K. G. (2010). The attrition crisis in South African universities: How to keep students on the graduation path. *Journal of Interdisciplinary Economics, 22*(3), 181–201.

Cohn, E., Cohn, S., Balch, D. C., & 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., & Chan, A. (2012). High school grades and university performance: A case study. *Economics of Education Review, 31*, 524–542.

Di Pietro, G. (2004). The determinants of university dropout in Italy: A bivariate probability model with sample selection. *Applied Economics Letters, 11*(3), 187–191.

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

Fredriksson, P., Öckert, B., & Oosterbeek, H. (2014). *Inside the black box of class size: Mechanisms, behavioral responses, and social background*. IZA Discussion Paper No. 8019, Institute for the Study of Labor, Bonn, Germany.

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

Gury, N. (2011). Dropping out of higher education in France: A micro-economic approach using survival analysis. *Education Economics, 19*(1), 51–64.

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.

Johnes, G., & McNabb, R. (2004). Never give up on the good times: Student attrition in the UK. *Oxford Bulletin of Economics and Statistics, 66*(1), 23–47.

Kerkvliet, J., & 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. (2003). Economic considerations and class size. *Economic Journal, 113*, F34–F63.

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

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

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.

Rask, K. (2010). Attrition in STEM fields at a liberal arts college: The importance of grades and pre-collegiate preferences. *Economics of Education Review, 29*, 892–900.

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

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

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.

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

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

Vaithianathan, R., Maloney, T., Putnam-Hornstein, E., & 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.

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