Shifting from Retention Rates to Retention Risk: An Alternative Approach for Managing Institutional Student Retention Performance

Prepared by Mark Conrad and Katherine Morris, York University
for the Higher Education Quality Council of Ontario

Higher Education
Quality Council
of Ontario

An agency of the Government of Ontario

Disclaimer:

The opinions expressed in this research document are those of the authors and do not necessarily represent the views or official polices of the Higher Education Quality Council of Ontario or other agencies or organizations that may have provided support, financial or otherwise, for this project.

Cite this publication in the following format:

Conrad, M and Morris, K. (2010). *Shifting from Retention Rates to Retention Risk: An Alternative Approach for Managing Institutional Student Retention Performance.* Toronto: Higher Education Quality Council of Ontario.

Published by:

The Higher Education Quality Council of Ontario

1 Yonge Street, Suite 2402
Toronto, ON Canada
M5E 1E5
Phone: (416) 212-3893
Fax: (416) 212-3899
Web: www.heqco.ca
E-mail: info@heqco.ca

Table of Contents

Table of Contents

Executive Summary

In the past, the term “persistence” was used somewhat interchangeably with “retention” to describe the *fact* of students *remaining* in a course of studies from one year to the next, typically at a single institution and sometimes within a particular program. Over the last few years, however, persistence has shifted in meaning to refer to the *ability* of students to *continue* their PSE studies and ultimately graduate, regardless of switches between programs or institutions or even temporary absences from PSE altogether. There is a growing recognition in Ontario and across Canada that this system-wide perspective on persistence will help government and institutions manage a highly functional, well-integrated PSE system, one in which students can avail themselves of numerous alternative educational opportunities and pathways to success.

It would be a mistake, however, to assume that these system-wide concerns are the primary arena in which PSE outcomes ought to be managed. Indeed, the concept of persistence as a process whereby students overcome obstacles is of note only in the context of the presence of initial decisions to leave and not return to a particular institution. The central aim of any university ought to be to improve its own retention of students. Indeed, a sustained focus on improving *in situ* retention outcomes is a vital component of an overall strategy for achieving high system-wide persistence rates. It is in the best interests of government and universities to develop the means by which retention practice efficacy can be reliably assessed, compared amongst institutions and used within institutions to actively improve retention rates.

Unfortunately, two common approaches used to calculate retention rates – the raw rate approach and the natural rate approach – are seriously flawed and cannot be recommended for use by Ontario PSE institutions as tools for managing retention practices.

The raw rate approach is transparently inadequate. The crux of the problem with raw rates is that they are essentially outcome measures unadjusted for variation in inputs. An institution that is in a position to admit students who are highly prepared academically, financially and culturally for university life at that particular institution can expect to be rewarded with relatively high outcome rates, and this without having to innovate or invest much in retention practices. Evaluating retention practice efficacy on the basis of raw rates favours institutions that are able to offload potential retention risks during the admissions process.

Another common approach used to calculate retention rates is to calculate the differences between raw rates and “expected” or “natural” rates and then to base evaluations and comparisons on these differences. Natural institutional rates are averages of the estimated probabilities of an event occurring (e.g., being retained after one year, graduating within four years) for each member of a cohort of students at an institution. One key feature of the statistical models upon which the probability estimates are based is the fact that they are system-wide models, pooling data across all institutions in the study and delivering a single set of model coefficients that is applied to all institutions. Another key feature is the fact that probability estimates are based on predictor variables that usually include only pre-entry characteristics of students and sometimes include environmental characteristics such as institution size, the field of study and whether the school primarily serves urban commuters. An institution with a raw rate that exceeds its natural rate is deemed to be performing well at retaining students, whereas an institution with a raw rate that is lower than its natural rate is evaluated as performing poorly. This approach has been implemented in the United States but not in Canada.

Three interpretation problems are ingrained in the natural rate approach that impede its meaningful application: normative interpretations given to natural rates are unwarranted; attributions of causation – to students in the case of natural rates and to institutions in the case of differences between natural and raw rates – are also unwarranted and potentially misleading; and a single set of system-wide coefficients is not likely to provide useful characterizations of the realities in play at individual institutions. A large and growing body of research embeds retention processes within the local context of individual institutions and indeed individual students. As research findings accumulate, there is a deeper and growing appreciation of the fact that the PSE system is not homogeneous in terms of the magnitude or direction of relationships between factors influencing retention event occurrence and the actual occurrence of those events. Rather, processes generating retention events operate locally and with considerable variation in form and intensity amongst locales, so system-wide characterizations do not give meaningful summaries of local conditions. The natural rate approach looks like a more sophisticated, finely tuned analysis, but its looks are deceiving.

An alternative to the raw and natural rate approaches is to move away from retrospective analyses of retention rates in favour of prospective analyses of retention risks. According to this approach, institutions use historical data to develop statistical models of retention risk at the individual student level. These models are then employed to estimate for each student in a currently enrolled cohort the “risk” (expressed as a probability) of continuing with their studies beyond a certain length of time.

The models used in the analysis must include a wide array of predictor variables, not just those pertaining to the pre-entry characteristics of students. The goal of the modelling exercise is to produce an accurate estimate of the actual retention risk faced by each individual student. These probabilities can be rolled up to produce an estimate of institutional-level risk exposure. These institutional risk forecasts differ from the natural rates discussed above in that they are estimates of the risk of events that have not yet occurred (natural rates are retrospective) and in that they are intended to accurately estimate full actual risk as opposed to only the portion of risk that is “due to the student.” No normative or causative assumptions are required by the logic underpinning the interpretation of the risk estimates. With the passage of time, retention outcomes become known and the efficacy of institutional retention practices are evaluated by comparing actual outcomes with the previously estimated risks.

There are numerous benefits to this approach. First, the three problems associated with the interpretation of natural rates are circumvented. The approach focuses on retention outcomes that have not yet occurred, and it provides student-specific, prospective assessments of risk with regard to the occurrence of these outcomes. Thus, the approach provides highly actionable information that is easily integrated into an institution’s practices for managing retention. Data on the participation of individual students in various retention initiatives can be incorporated into the analysis, either at the risk estimation stage or during subsequent evaluations of retention initiatives, or both, as appropriate. Given that the risk estimates are prospective in nature, they may be used as inputs to other analyses, such as enrolment forecasts. Finally, given that risk assessments are expressed as probabilities, with values that always exist in the range of 0 to 1 and have consistent interpretations, it ought to be possible to directly compare risk assessments across models and thus across institutions.

We explored this alternative approach in a pilot study using undergraduate enrolment data from York University, located in Toronto, Ontario. In our particular implementation of the approach, we used an extension of classification and regression trees to survival analysis – specifically a "random survival forest" (RSF) algorithm – to analyze time-to-stop-out data for 83,593 students who entered York University as new, direct-entry, first-year undergraduates from fall 1996 through fall 2006. The term "stop-out" is defined here as a failure to re-enroll at an institution, either temporarily or permanently. Thus, stop-out is used as an inclusive term encompassing temporary and permanent leavers from an institution without regard for students' eventual status, if any, at other PSE institutions. The main output of the analysis is a set of survival probability estimates for each individual student, given at 1-year intervals (i.e., probability of the student surviving [continuing] beyond 1 academic year, probability of surviving beyond 2 academic years and so on). The survival probabilities are interpretable as retention risk estimates, and the complements of the survival probabilities are interpretable as stop-out risk estimates. Survival probability estimates were used as the basis for binary classifications of students into "stop-outs" and "continuers" and were also evaluated in terms of their accuracy and thus usefulness for estimating institutional exposure to retention risk.

We were able to achieve a single predictive classification of stop-outs that correctly identified slightly more than 90 percent of first-year stop-outs, more than 35 percent of second-year stop-outs, and that had a predictive specificity of 82 percent (i.e., false positive rate of 18 percent).

Survival probability (i.e. retention risk) estimates for those students most at risk of stopping-out by the end of first year were higher than actual survival rates (i.e., retention rates) and therefore not suitable as the basis for assessing institutional stop-out risk exposure or evaluating the efficacy of retention interventions. It is felt that a more sophisticated data set would likely produce more accurate estimates, which would then be suitable for risk exposure assessments and intervention evaluations. To that end, future work ought to include developing "early warning" data focusing on initial conditions faced by students as they begin an academic year and "in-stream" data that becomes available as an academic year progresses. Early warning data might provide indications of students' goals, commitments, preparedness and involvement in their studies and indicators of outside demands on their time and attention, as well as correlates of social and cultural capital and even measures of emotional intelligence (which has been related to adaptability to university life). In-stream data might include early data regarding academic performance in "gatekeeper" courses, indicators of integration and engagement, and even information regarding students' stated intentions to persist in their studies.

A program of interview- and/or focus-group-based qualitative research that delves into the university experiences of those who leave their studies would help institutional researchers prioritize the various options available to them as they seek to create more sophisticated retention risk prediction data sets. Indeed, the current gap between available data and what is required to produce precise risk estimates is an issue to be grappled with. The advantages of a prospective, retention-risk-based approach are many, but there is substantial work to accomplish in order to make the approach fully viable.

Why a Shift in Approach Is Desirable

System-Wide Persistence and Institution-Level Retention

In their review of research on retention and attrition at universities and colleges, Grayson and Grayson (2003) observe that examining retention and attrition solely from the perspective of institutional rates may "paint a misleading picture," particularly if one's goal is to understand system-wide attrition and retention. This is because many students who drop out of one institution continue their postsecondary education (PSE) at a later date and often at a different institution. In the years since the publication of the review, a number of studies have been published that use longitudinal data sets to track the cross-institutional, system-wide enrolment of individual students (see Parkin and Baldwin, 2009, for a recent review).

Aside from providing a better picture of system-wide PSE attrition rates – five-year drop-out rates of 10 percent for university students versus 26 percent when single-institution data sets are used (Finnie and Qui, 2008) – these more recent studies have helped entrench a meaning of the term "persistence" that is quite distinct from the meaning of the term "retention." In the past, persistence and retention were used somewhat interchangeably to denote the *fact* that a student *remained* in a course of studies from one year to the next, typically at a single institution, and sometimes within a particular program. Over the last few years, however, the term "persistence" has shifted in meaning to refer to the *ability* of a student to *continue* PSE studies and ultimately graduate, regardless of switches between programs or institutions or even temporary absences from PSE altogether. Persistence highlights the fact that, these days, students move through PSE in various ways and some students seem able to make adjustments that lead to a continuation of studies, whereas other students seem less able to do so. Persistence thus appears to be related to resilience, which is "the capacity to overcome obstacles, adapt to change, recover from trauma or to survive and thrive despite adversity" (Canadian Career Development Foundation, 2007; Parkin and Baldwin, 2009).

An implication of this new perspective is that mobility of students within the PSE system is a fact of life to be acknowledged and dealt with by governments and PSE institutions. One way to deal effectively with the fact of mobile students is to create a PSE system that allows students to select from numerous alternative educational opportunities and to do so at various stages of their PSE experience, not just the beginning. Initiatives that support students' resiliency or lower the level of resiliency required of students (by lessening unnecessary obstacles and adversity associated with changing course while pursuing PSE studies) would also help.

It would be a mistake, however, to assume that these system-wide concerns are the primary arena in which retention/attrition/persistence outcomes ought to be managed. Indeed, persistence as a system-level concern is of note only in the context of initial decisions to leave and not return to a particular institution. The first line of defense in achieving high system-wide persistence rates must be for institutions to help students achieve their PSE goals *in situ*. As

Parkin and Baldwin (2009) assert, PSE institutions must identify which of their students are at elevated risk of leaving their studies and "provide them with support programs created for and tailored to them so that they can make the necessary adjustments over time and succeed." Parkin and Baldwin go on to say that "institutions will increasingly need to focus on the 'micro' level of subsets of their student populations. Their actions regarding these groups will help determine the success of the Canadian post-secondary system as a whole." That is to say, while some stop-out decisions are in no way a failure on the part of the institution or student, research suggests that many others are indeed failures of a kind.

Whether financing, academic preparedness, social and cultural capital deficits or other factors play a role, identifiable root causes do lead to many stop-out decisions. It is surely not enough to say that these situations are adequately dealt with solely by easing the selection of alternative PSE pathways. That response would simply create a pool of students who are persistent wanderers, settling for educational experiences not of their liking but within their grasp, given their current financial, cultural and personal resources. Rather, individual institutions must strive to understand their students and understand themselves as institutions and the ways in which they interact with students. Furthermore, governments ought to recognize that a continued focus on improving *in situ* retention outcomes is a vital component of an overall strategy for achieving high persistence rates.

Within this context, a second observation in Grayson and Grayson's (2003) report is worth further attention. After considering the research on who leaves their studies and why, they state the following:

> [I]t could be very misleading to make general statements about who drops out and why. In some situations, factors like grade point average contribute to persistence; in others they can have no impact, or negative impact, on persistence. The only factor that probably has a consistent relationship to retention is the expressed intent of students to continue their studies in the coming year. There is also some reason to believe that, in commuter institutions, academic integration is more important in explaining persistence than in residential colleges and universities. This does not mean that we should abandon the possibility of understanding student persistence and attrition. It does mean that for the time being we should recognize that the explanations we have are institution specific . . . .In different institutional settings different factors explain attrition.

As will be discussed more fully in the next section of this paper, "who leaves and why" increasingly appears to be a question best answered within the local setting of particular institutions and their students. Detailed yet generally applicable (e.g., system-wide) answers to the question simply may not be realistic, and studies that do not take institution-specific effects into account can be challenging to interpret appropriately and apply correctly.

To summarize, the argument being made here is that in spite of the importance of understanding PSE system-level persistence amongst Ontario students – and implementing policies that do not unduly impede system-level, cross-institutional persistence – the central aim of any institution ought to be to improve its own retention of students. Furthermore, it is in the best interests of government and institutions to develop the means by which retention risks can be reliably assessed, compared amongst institutions and used within institutions to actively improve retention rates.

Institutional Retention Rates: The Raw Rates Approach

With the passage of the federal Student Right to Know and Campus Security Act (Public Law 101-542) in 1990, American colleges and universities were compelled to publish institution-specific academic outcome statistics (i.e., completion and graduation rates). Astin (1993b, 1997) makes the point that the Act's language implies an underlying assumption that knowing these rates would allow students to make informed decisions regarding which college or university to attend because the higher the graduation or completion rate, the better job an institution is doing at retaining and subsequently graduating its students. Institutions with lower rates are assumed to be doing a relatively poor job. In other words, from the outset, raw rates (also called absolute, simple or actual retention rates) have been used to gauge institutional performance and effectiveness via cross-institutional comparisons. The publication of raw retention and retention-related rates (e.g., attrition, completion, graduation) is now commonplace in the United States, Canada and elsewhere, as are the subsequent cross-institutional comparisons.

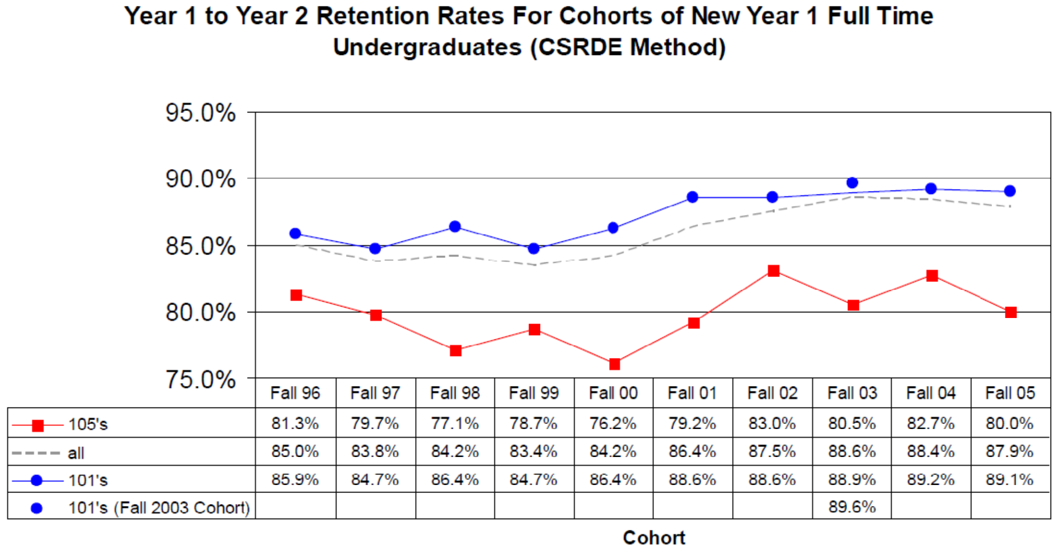
Yet, using raw rates to compare the effectiveness and performance of institutions is ill conceived. Astin (1993a, 1993b, 1997) noted that raw, rate-based institution comparisons of degree completion rates are hopelessly confounded by factors related to who is admitted to the institutions in the first place. His longitudinal analysis of data from 39,243 students attending 129 four-year colleges and universities showed that over half of the variation amongst institutions could be statistically explained by pre-entry characteristics of the students themselves, without any reference in the statistical models to the influence that the institutions might have had on academic outcomes (Astin, 1993b). Similarly, Astin and Oseguera (2005) conducted a longitudinal study using data from the 1994 CIRP Freshman Survey and found that two-thirds of the variation amongst institutions in terms of degree completion rates was statistically explained by the pre-entry characteristics of those admitted. Astin (1997) also noted the importance of environmental factors such as field of study, institution size and whether an institution serves mainly urban commuter students or mainly residential students. He was careful, however, to highlight the fact that the effect of these environmental factors on raw rates exhibits considerable variation. The central conclusion from these studies is that it is unwise and misleading to compare the raw rates of degree completion (and by extension raw retention or attrition rates) of different institutions without taking into account the pre-entry characteristics of those admitted to the institutions in the first place.

The basic problem in the analysis of raw rates is that they are essentially outcome measures unadjusted for variation in inputs and thus are rather transparently inadequate. An institution that is in a position to admit only students who appear to be the most prepared academically, financially and culturally for university life at that particular institution can expect to be rewarded with relatively high outcome rates, and this without having to invest much in effective retention programs. In fact, the most selective of institutions could conceivably (and if so inclined) ignore the needs of their higher-risk students and still exhibit excellent retention rates overall. Such institutions effectively solve their retention problem by not admitting higher-risk students to begin with.

While this is a legitimate admissions strategy for some institutions, it is not feasible for all schools, and its broad application across the PSE system would certainly run contrary to public policy objectives, according to which PSE opportunities need to be made widely available to prospective students across the spectrum of social, economic, cultural and (within reason) academic backgrounds. At the other end of the spectrum are universities whose admission policies, socio-cultural context and perhaps even mandate contribute to the achievement of lower raw retention rates. These institutions may strive to continuously improve their retention programs and may even enjoy considerable success in the undertaking, yet never hope to achieve the raw rates of institutions that are not exposed to the same retention risks to begin with. And when governments and prospective students use raw rates as a means of evaluating the relative efficacy and performance of institutional retention practices, those institutions whose risk exposure is high also run the risk of not having their retention efforts fairly recognized. In short, evaluating retention program efficacy on the basis of raw rates rewards institutions that are able to offload potential retention risks during the admissions process.

Furthermore, raw rates for a particular institution may change over time in response to shifts in the characteristics of its student body and not due to changes in retention practices. For example, in reporting first- to second-year retention rates to the Ontario government as part of its 2006 Multi-year Accountability Action Plan “report-back,” York University noted that year-over-year retention rate dynamics appeared sensitive to the proportion of students enrolling directly from Ontario high schools (York University, 2007). (These students are commonly referred to as “101s” for administrative coding reasons.) The non-101 group (called “105s”) typically exhibits retention rates that are lower and much more dynamic than the rates for 101s (see Figure 1). This leads to situations such as the one that occurred with the cohort of students entering in 2003. That cohort consisted of a higher proportion of 101s than usual, and thus 101s controlled the overall retention rate (the dashed line in the figure) in spite of a large drop in the retention rate of non-101s.

Figure 1
Retention rates for entering-year cohorts of students (new, year 1, full-time students) entering York University from fall 1996 through fall 2005. Rates are calculated according to the method used by the Consortium for Student Retention Data Exchange (CSRDE)


Year 1 to Year 2 Retention Rates For Cohorts of New Year 1 Full Time Undergraduates (CSRDE Method)

	Fall 96	Fall 97	Fall 98	Fall 99	Fall 00	Fall 01	Fall 02	Fall 03	Fall 04	Fall 05
🟥 105's	81.3%	79.7%	77.1%	78.7%	76.2%	79.2%	83.0%	80.5%	82.7%	80.0%
- - - - all	85.0%	83.8%	84.2%	83.4%	84.2%	86.4%	87.5%	88.6%	88.4%	87.9%
● 101's	85.9%	84.7%	86.4%	84.7%	86.4%	88.6%	88.6%	88.9%	89.2%	89.1%
● 101's (Fall 2003 Cohort)								89.6%		

Source: York University, 2007.

Institutional Retention Rates: The Natural Rates Approach

An alternative to using raw rates is to calculate the differences between raw rates and “expected,” or “natural,” rates and then to base evaluations and comparisons on these differences. Natural institutional rates are averages of the estimated probabilities of an event occurring (e.g., being retained after one year, graduating within four years) for each member of a cohort of students at an institution. Key features of the statistical models upon which the probability estimates are based include (1) the fact that they are system-wide models, pooling data across all institutions in the study and delivering a single set of model coefficients that is applied to all institutions, and (2) the fact that probabilities are estimated based on predictor variables that include pre-entry characteristics of students and sometimes environmental characteristics such as the field of study, institution size and whether the school serves primarily urban commuters (Astin, 1993a, 1993b, 1997; Higher Education Research Institute, 2003). In the case of retention rates, an institution with a raw rate that exceeds its natural rate is deemed to be performing well at retaining students, whereas an institution with a raw rate that is lower than its natural rate is evaluated as performing poorly. Proponents of the use of natural rates have argued that the procedure circumvents the pitfalls of the raw-rate-only analysis and instead provides an “internal standard” (the natural retention rate) against which an institution can be judged. In effect, argues Astin (1993b), the institution is being “compared with itself.” The degree to which the institution is over- or underachieving relative to rates expected of it is the metric used to compare institutions.

The natural rate approach appears to hold merit over the transparently inadequate raw-rate-only approach, and it has been operationalized several times. Natural graduation rate formulae for American colleges and universities were published by Astin (1997) and Astin and Oseguera (2005), and since 1997, *US News and World Report* has published "graduate rate performance" measures, which are essentially differences between raw and natural graduation rates. No natural rate formulae have been developed for Canadian institutions, a situation considered "unfortunate" by Grayson and Grayson (2003).

There are, however, three interpretation problems ingrained in the natural rate approach that impede its meaningful application: normative interpretations given to natural rates are unwarranted and potentially harmful; attributions of causation – to students in the case of natural rates and to institutions in the case of differences between natural and raw rates – are unwarranted and misleading; and a single set of system-wide coefficients is not likely to provide useful characterizations of the realities in play at individual institutions.

A normative interpretation is routinely applied to natural rates. Raw rates that fall to one side of the norm are taken to indicate deficiency or failure (institutions doing a poor job) and those that fall on the other side of the norm are viewed as indicating efficacy or excellence (institutions doing a good job). Analyses making use of the natural rate approach appear to accept uncritically this normative assumption. Readers are asked, "How good is your retention rate?" and are then advised to use the natural rate approach to find the answer (Astin, 1997; Higher Education Research Institute, 2003).

It is easy, however, to understand why such a normative assumption is unwarranted. First, there is nothing to stop the effect of shortcomings in retention practices from showing up partially in the natural rate – this will happen to the extent that the effects of shortcomings co-vary with pre-entry characteristics – and partially in deviations from the natural rate. Second, natural rates are merely statistical expectations, ones that change over time as the underlying data set evolves. There is no intrinsic desirability or sanction associated with a statistical expectation. It might be that a relatively poor natural rate is expected for a particular gender or ethnic group without that rate being desirable. So giving the natural rate a normatively neutral status seems inappropriate. The suitable interpretation of natural rates is thus non-normative: raw rates that fall to one side of the natural rate should simply be viewed as indicating rates that are worse than the system-wide expectation, and those that fall on the other side of the norm ought to be viewed as being better than the system-wide expectation.

Under no circumstances should one ignore the question of whether or not the expectation itself is good enough. It is easy to agree with the previous statement but apparently more difficult to remain faithful to it. Indeed, it is our opinion that the natural rate approach invites normative interpretation; even if one is careful not to frame interpretations explicitly in a normative manner, we believe there will always be a tendency for stakeholders to preoccupy themselves with institutional positioning relative to the natural rate and for institutions to feel a sense of accomplishment when "outperforming" the natural rate, regardless of what that natural rate might be suggesting. In a sense, all stakeholders – save for students themselves – are "let off the hook" for natural rates.

The natural rate approach also suffers an underlying rationale that requires casting students (and, to a much lesser extent, environment) as the causal agents generating natural rate values and institutions as the causal agents of the differences between raw and natural rates. The natural rate is taken as an “internal standard,” a product of the particular students admitted by the institution. Astin (1993b) suggests that the natural rate approach allows institutions to ask, “How well are we doing, given the students we admit?” Students are responsible for generating natural rates, and institutions are credited with moving these rates from their natural values to the actual raw values. Thus, the differences between natural and raw rates are taken to measure efficacy of retention practices. In other words, the relationship between students’ pre-entry characteristics and the academic outcome of interest (say retention after year 1) – that is to say, the relationship as described by the regression model used to generate natural rates – is assumed to be both structurally accurate and causal.

Yet the regression models merely describe empirically derived functional relationships using only pre-entry characteristics as predictor variables; the models’ structure and parameterization do not necessarily reflect the actual causal processes that generate academic outcomes (any more than hemlines actually cause recessions, for instance). And as previously mentioned, the effect of shortcomings in institutional retention practices (or the effects of any other factors not included in the model) can influence model parameterization to the extent that the effects of institutional shortcomings co-vary spuriously with students’ pre-entry characteristics. Furthermore, the assumed structural relationship between pre-entry characteristics and outcomes is regarded (implicitly) as sufficiently specified, and differences between natural and raw rates are attributed entirely and uncritically to the retention practices of institutions.

In fact, the problem with devising models that allocate causation to students on the one hand and institutions on the other hand runs much deeper than the concerns raised above. There is a growing body of research that highlights the importance of the myriad relations between the student, the institution and the wider socio-cultural environment. In these studies, the act of leaving one’s studies is cast as an outcome of the *interactions* between all involved, the implication being that it would be wrong-headed to attempt a simple decomposition of retention risk into a strictly student-caused portion and a second strictly institution-caused portion. These studies often refer to the “fit” between student and institution and invoke the concepts of social and cultural capital as a way of understanding the production of stop-out decisions. Perhaps the most obvious examples of this interactionist perspective are the many analyses that use Tinto’s theory of student integration (Tinto, 1975, 1993) or elaborations of it (Cabrera, Nora, & Castaneda, 1993; Sandler, 2000; Thomas, 2000) as a theoretical foundation.

According to Tinto’s theory, students exhibit various pre-entry characteristics and also related initial goals and expectations regarding their PSE studies and future career. The institution brings to the table its own expectations, goals and commitments. Student, institution and wider environment interact, producing student experiences, both social and academic. The salient aspect of these experiences is whether students integrate into the institution’s social and academic setting; decisions to leave or continue are based on students’ internal assessments of integration. If students feel a sense of congruence or good “fit” between their own (adjusted) intentions, goals and commitments on the one hand and their university experiences on the other hand, then they are likely to continue in their studies; otherwise, they are more likely to stop-out. A key point of this body of theory – one that is sometimes forgotten in its application – is that experiences, integration and retention events (i.e. either leaving or continuing) are generated by a dynamic system of interactions between student, institution and environment. Retention events are born in that nexus. One can speak of retention rates being “explained” by pre-entry characteristics, but this is true only in an empirical modelling sense. The reality of the situation is more complex, and this is likely why there is such wide variation in the size and even direction of the effects of specific pre-entry characteristics from one study to another and from institution to institution.

The importance of this point is given deeper consideration in some of the studies on first generation student (FGS) access to, and persistence in, PSE. The FGS concept has seen a variety of operational definitions, but in broad strokes, refers to students whose parents do not have [a specified level of] PSE experience. The definition is simple, but the underlying reasons why FGSs might experience university differently than non-FGSs are acknowledged to be fairly complex (see Auclair et al., 2008, for a recent review). In fact, it might be argued that the FGS concept is best viewed as a container or placeholder for a cluster of the many factors that generate students’ *habitus*. *Habitus*, as defined by Bourdieu (1977, 1990), is the set of durable and transposable skills and dispositions of an individual, including an individual’s schemes of perception, knowledge and thought, as well as dispositions to interpret experiences and act in certain ways. Durability refers to the idea that *habitus* changes slowly over time. But change it does, in response to an individual’s accumulation of experience as he/she negotiates through different fields.

A “field,” in the context of Bourdieu’s social theory, is defined as a social arena with its own set of norms and expectations, in which individuals contest for the possession of various kinds of capital. Universities offer numerous overlapping fields as students navigate through classrooms, clubs, pubs and dorms; participate in athletics; and visit the offices of the Registrar or the Student Aid unit – and the forms of capital in question are mainly social and cultural. It might be argued that the central problem for many FGSs is that their *habitus* does not translate readily into social and cultural capital according to the norms and expectations of the various fields of university life. According to this view, FGSs are more likely than non-FGSs to experience an acute disconnect between their *habitus* and those fields and to perceive large social and cultural capital deficits as a result. These deficits are interpreted as a “lack of fit,” which becomes a central feature of many FGSs’ perceived university experience. Looming capital deficits, not to mention the potential trauma of “shedding one social identity and taking on another” (London, 1996), make it difficult for these FGSs to acquire the additional social and cultural capital on offer at universities. In other words, “You’ve got to spend capital to make capital” or perhaps “The rich get richer and the poor don’t.”

Not all FGSs, of course, fail to adapt to university life (Ishitani, 2006; Lohfink and Paulsen, 2005). Indeed, *habitus* extends well beyond the FGS/non-FGS dichotomy to encompass all skills and dispositions, including those that contribute to adaptability, resilience and persistence (Duggan, 2002; Pascarella, Pierson, Wolniak, &Terenzini, , 2004). Furthermore, there are numerous other concepts in addition to FGS that also serve as organizing ideas for considerations of retention vis-à-vis *habitus*. Emotional intelligence is one example and another is recent immigrant status. The latter will grow in importance for universities in the Greater

Toronto Area (GTA) as recent immigrants continue to represent larger proportions of the GTA population.

Bourdieu's concepts of *habitus*, fields, cultural and social capital, and lack of fit – and the social theory knitting them together – are useful in particular because they help researchers delve deeply into the idea that retention events are generated in the continuous stream of interactions between student and institution. Moreover, these concepts position institutions not as agents, but as locales of numerous interrelated social fields, each a social system of many interacting people, any one of whom might make the difference in a particular student's evaluation of university experiences and subsequent decision to leave or continue studying. Official institutional policies and procedures establish a portion of the norms and expectations faced by students, but even these official norms and procedures are experienced by students mainly via their interactions with other people – students, faculty and staff – each with their own *habitus*. The net effect of these interactions may be surprising and difficult to predict. Lehmann (2007) notes, for instance, that even subtle differences in the characteristics of the student population seemed to have profound effects on the experiences perceived by the students who participated in his study. Lehmann's research also supports an idea familiar to many PSE enrolment managers: students tend to evaluate university experiences, including perceived social and cultural capital deficits, in relative terms against competing social, cultural and indeed financial capital wins and losses expected in numerous non-university fields.

So when it comes to factors generating retention events (leaving or continuing), the devil really is in the details of each student's experience. Generalizations are possible, of course, but usefully accurate retention models – the kind needed to make meaningful statements about how well institutions are doing at retaining students – are likely to require a heavy dose of interaction effects, nonlinearities and student-level data. Developing simple models that accurately partition out a student-caused portion of retention rates based on a limited set of pre-entry characteristics increasingly seems like an unrealistic project.

Finally, the natural rate approach relies on pooled data from students at many institutions and the estimation of a single set of system-wide model coefficients. Data for individual students are then input into a formula parameterized by these system-wide coefficients in order to calculate each student's expected probability of experiencing the academic outcome of interest. Natural rates for a particular institution are then calculated as the average of the expected probabilities for students attending that institution. This approach is valid if the single set of system-wide coefficients does an adequate job of reflecting conditions that really occur at each institution. If not, then an ecological fallacy is being committed, in which erroneous inferences about the nature of objects and relations at a lower level, the institution level, are based on statistics and relationships that hold at an aggregate level, in this case, the system-wide level).

Unfortunately, Grayson and Grayson's admonition against making general (e.g., system-wide) statements about who drops out and why remains as relevant today as it was in 2003. Not only are we in a poor position to make relevant system-wide statements regarding who leaves and why, but it seems ever more likely that we will never be in a position to make such statements because processes generating decisions to leave or continue with studies (which roll up into retention rates) operate at a much more local level. As research findings accumulate, a deep appreciation is developing of the fact that the PSE system is not homogeneous in terms of the magnitude or direction of relationships between factors influencing retention event occurrence and the actual occurrence of those events. Rather, processes generating retention events operate locally and with considerable variation in form and intensity amongst locales; system-wide characterizations do not give meaningful summaries of local conditions. The natural rate approach looks like a more sophisticated, finely tuned analysis, but its looks are deceiving.

An Alternative Approach: Shifting from Retention Rates to Retention Risk

An alternative to the natural rate approach is for institutions to use historical data to develop statistical models at the student level, and then use these models to estimate, for each individual student in a currently enrolled cohort, the risk (expressed in terms of a probability) of being retained beyond some point in the future. The models used in the analysis should include a wide array of predictor variables, not just those pertaining to the pre-entry characteristics of students. The objective of the modelling exercise is to produce an accurate estimate of the actual retention risk faced by each student at that particular institution. These probabilities can be rolled up in some way (e.g., FTE weighted average) to produce an estimate of institution-level retention risk exposure. These institutional risk forecasts differ from the natural rates discussed above in that they are estimates of the risk of events that have not yet occurred (natural rates are retrospective) and in that they are intended to accurately estimate full, actual risk as opposed to only the portion of risk that is “due to the student.” No normative or causative assumptions are required (or invited) by the logic underpinning the interpretation of the risk estimates. With the passage of time, retention outcomes become known, and the efficacy of institutional retention practices are evaluated by comparing actual outcomes with the previously estimated risks.

There are numerous benefits to this approach. First of all, the three problems discussed in the previous section, which are associated with the interpretation of natural rates, are circumvented. The approach focuses attention on academic outcomes that have not yet occurred and provides student-specific, prospective assessments of risk with regard to the occurrence of these outcomes. Thus, the approach provides highly actionable information that can be integrated into an institution’s practices for managing retention. Data on the participation of individual students in various retention initiatives can be incorporated into the analysis, either at the risk estimation stage or at the intervention evaluation stage, or both, as appropriate. Given that the risk estimates are prospective in nature, they may be used as inputs to other analyses, such as enrolment forecasts. Finally, given that risk assessments are expressed as probabilities, it ought to be possible to directly compare risk assessments across models and thus across institutions.

A few additional comments are worth making. Ideally, the statistical modelling technique used to develop the risk assessments ought to be able to handle the inclusion of many predictor variables, as well as interaction effects, collinearities and nonlinearities. Also, when selecting a modelling technique, it should be noted that prediction performance is key to the success of the overall approach. Finally, accurate assessments of retention risks are required as early on as possible, in order to leave enough time for retention interventions.. This means that early warning data, such as engagement and preparedness survey results and early academic results, are required.

We explored this alternative approach in a pilot study using enrolment data from York University in Toronto, Ontario, Canada, to estimate retention risks and to predict stop-out events for undergraduate students.

Pilot Study Overview: Estimating Stop-Out Risks and Predicting Stop-Out Events at York University

In our particular implementation of the approach suggested above, we used an extension of ensemble classification and regression tree techniques to survival analysis – specifically a “random survival forest” (RSF) algorithm (Ishwaran and Kogalur, 2007; Ishwaran, Kogalur, Blackstone, & Lauer, 2008) – to analyze time-to-stop-out data for students who entered York University as new, direct-entry, first-year undergraduates from fall 1996 through fall 2006. Using survival analysis techniques to study retention risks is not new (DesJardins, McCall, Ahlburg, & Moye, 2002; Ishitani & DesJardins, 2003; Ishitani & Snider, 2006; Radcliffe, Huesman, Kellogg, & Jones-White, 2009). More unusual is using an ensemble tree technique to accomplish the survival analysis, yet there appear to be a number of benefits to doing so, including the fact that these techniques perform extremely well in many cases (Berk, 2006; Breiman, 2001). They are also quite flexible in dealing with nonlinearities and other aspects of large, complex data sets (Ishwaran, Kogalur, Blackstone, & Lauer, 2008), including the routine handling of multiple sub-classes within an overall classification. The term “stop-out” is defined in this study as a failure to re-enroll at the university (as of the official November 1 count date) either temporarily or permanently and without regard for eventual status, if any, at other PSE institutions. Stop-outs include “switchers,” who at some point continue studies in the PSE system – albeit at other institutions – and “permanent leavers,” who do not continue their studies. In the context of an institution attempting to reduce the number of students who leave in the first place, the key event is the stop-out, regardless of what students subsequently decide to do.

The main output of the analysis is a set of survival probability function estimates, a separate function being estimated for each individual student. Each of these functions consists of a set of survival probability estimates for each student, given at 1-year intervals (i.e., probability of surviving [continuing] beyond 1 academic year, probability of surviving beyond 2 academic years and so on). The survival probabilities are interpretable as retention risk estimates, and the complements of the survival probabilities are interpretable as stop-out risk estimates. Survival probability estimates were visually compared against actual retention rates to assess their accuracy and thus suitability for assessing institutional exposures to retention risk. The estimates were also used as the basis for binary classifications of students into “stop-outs” and “continuers”: students with survival probability estimates falling below a certain threshold were predicted to stop-out while those with estimates above the threshold were predicted to continue with their studies. The prediction performance of these classifications was assessed using confusion matrices (contingency tables of predictions versus observed actuals). A summary of the main results of the pilot study is given below, and details of the study are provided in Appendix A.

Estimated probabilities of survival past two years were used to predict stop-outs regardless of the year in which the stop-outs actually occurred (see Figure 2), for students in a test data set. (The test data were not used to build the predictive models, and this allowed for unbiased assessments of the model's prediction power.) Using a threshold value of 0.67, we were able to predict correctly 1,458, or 90.3 percent, of first-year stop-outs in the test data set (represented by red dots in Figure 2) and 363, or 35.3 percent, of second-year stop-outs (represented by green dots). Overall prediction specificity of the classification is 82.4 percent (i.e. an overall false positive rate of 17.6 percent; cf. Table 1).

Figure 2

Ensemble estimates of the probability of surviving beyond 2 academic years, by stop-out status. Red points indicate stop-out events occurring during first year, green points indicate stop-out events occurring during second year, yellow points indicate stop-outs occurring at some point after second year and blue points indicate students who continue uninterrupted to graduation. The red line represents a survival probability threshold-value of 0.67, below which students are predicted as stop-outs. Observations censored due to students still being enrolled (and not graduating) at the end of the study period are excluded from the figure, since the actual status for these students at the beginning of fall 2007 (i.e., the year immediately following the end of the study period) are unknown. Random noise has been added to the location of points along the x-axis to reduce over-plotting.

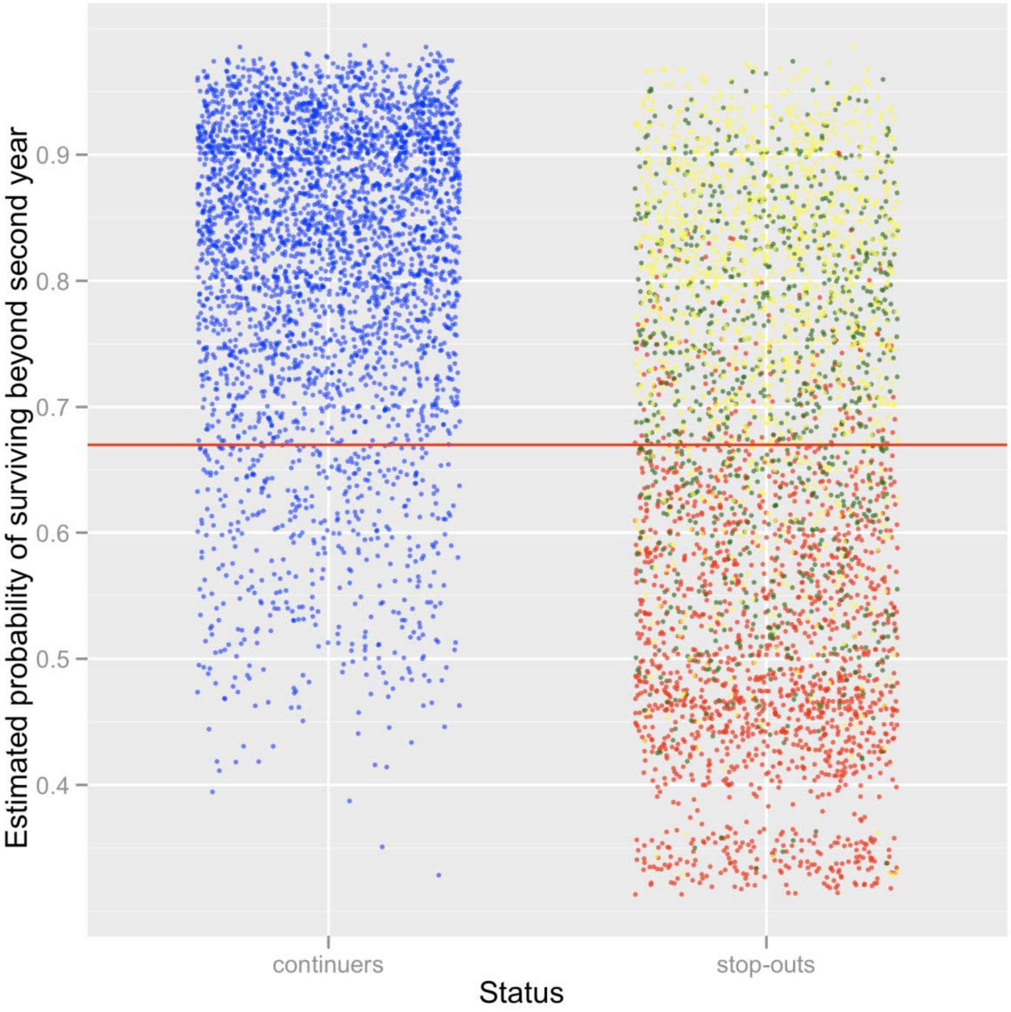


Table 1
Confusion matrix for a classification of students into two groups: “stop-outs” and “continuers.” The classification is based on a threshold value of 0.67 applied to ensemble estimates of the probability of surviving beyond 2 years, for students included in the test data set. Observations censored due to students still being enrolled (and not graduating) at the end of the study period are excluded from the confusion matrix, since the actual status for these students at the beginning of fall 2007 (i.e., the year immediately following the end of the study period) are unknown.

		Predicted		
		Continuers	Stop-outs	Total
Observed	Continuers	3,127	428	3,555
	Stop-outs	1,808	1,998	3,806
	Total	4,935	2,426	7,361

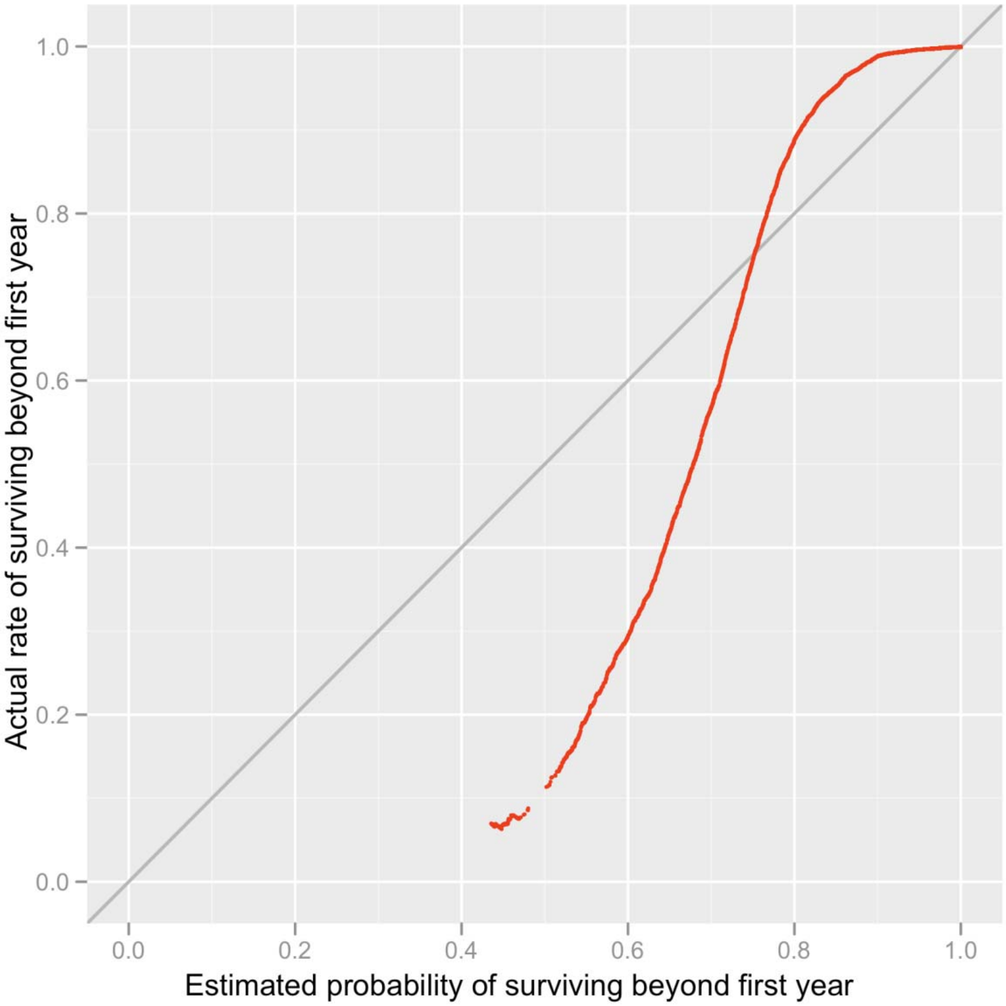
Thus, even with the limitations of the data employed in this pilot study (e.g., time measured in academic years instead of four-month terms or even individual months; few predictor variables relating to social, cultural and personal preparedness and demands; and no early academic progress data), an effective classification of first-year stop-outs was achieved. This result suggests that retention management personnel may look to this sort of prospective retention risk approach for actionable information, particularly since limited resources for retention management tend to become focused on attending to first-year students at risk of stopping-out, due to the fact that more stop-outs occur prior to the second academic year than during any other individual year.

The classification results for students at risk of stopping-out during second year or later, while still useful, are certainly less satisfying than the first-year stop-out results, particularly since many additional students stop-out during their second year of studies. It seems reasonable to expect, however, that the inclusion of predictor variables regarding academic progress, ongoing personal and financial demands, and even intentions to persist (or not) with studies would improve upper-year stop-out predictions.

A more sophisticated data set might also be expected to generate more accurate survival probability estimates, particularly for those students most at risk of stopping-out. Certainly, the survival probability estimates obtained in this pilot study are not reliable enough to use as the basis for an evaluation of institutional stop-out risk exposure. Figure 3 illustrates the relationship between estimated probabilities of surviving (i.e., continuing) past the first year of studies and actual rates of survival past the first year for students in the test data set. Ideally, estimated probabilities and actual rates would be nearly equal. In a general sense, the survival probability estimates are well behaved: higher probability values are associated with higher actual rates – an important result, since this is what allowed for an effective prediction of first-year stop-outs.

On the other hand, Figure 3 clearly shows that the estimated survival probabilities do not *closely* match actual rates except in two small regions of the plot (one of these regions is centered on the value 0.75 and the other near the value 1.0). In particular, students most at risk of stopping-out (those with low survival probability estimates) exhibit actual survival rates that are much lower than the probability estimates. In other words, stop-out risks for these students have been underestimated. Greater accuracy would be necessary in order to usefully employ the survival probability estimates as inputs to enrolment forecast models, institutional retention risk exposure assessments or retention practice evaluations. Future work subsequent to this pilot study will determine whether sufficient gains in accuracy can be realized through the use of more sophisticated data sets.

Figure 3
Plot of actual, localized rates of survival past the first year of studies, by estimated probabilities of surviving past the first year of studies. The black line in the plot references equality between actual rates and estimated probabilities.



In particular, future work ought to focus on achieving sufficiently accurate survival probability estimates after the first academic year, with the recognition that some of this work may also serve to improve estimates of survival after two or more academic years. Obtaining more accurate estimates will likely require the development of “early warning” data focusing on initial conditions faced by students as they begin an academic year and “in-stream” data that become available as an academic year progresses. Initial conditions data might provide indications of students’ goals, commitments, preparedness and involvement in their studies, and indicators of outside demands on their time and attention, as well as correlates of social and cultural capital and even measures of emotional intelligence (which has been related to persistence). In-stream data might include early data regarding academic performance in “gatekeeper” courses, indicators of integration and engagement, and even information regarding students’ stated intentions to persist in their studies.

To help prioritize the various options available to institutional researchers as they seek to create more sophisticated retention risk prediction data sets, an initial program of interview- and focus-group-based qualitative research delving into the university experiences of those who stop-out is suggested.

The ability of a single algorithm to generate risk assessments across a range of time intervals is one of the key strengths recommending the use of survival analysis techniques to analyze retention risks. (The other primary strength of the technique is that recent, albeit censored, data can be included in the analysis.) From a practical point of view, a survival analysis requiring only one input data set and providing retention risk estimates for all students, regardless of the year they are in, is far easier to handle than an analysis approach in which a separate data set and separate analysis run are required for each cohort of students. So although the prospective retention risk approach outlined in this report does not absolutely demand the use of a survival analysis technique, it would benefit from the use of one.

Returning to the matter of assessing the efficacy of institutional retention practices this report has outlined an approach based on prospective, student-level retention risk estimates compared against actual student-level retention outcomes once the data on actual outcomes becomes available. This approach offers several advantages: It does not require any normative or causative interpretations of the empirically derived survival quantities; it does not require a system-wide retention model, thereby reducing chances of committing an ecological fallacy (and making fallacious comparisons between institutions); and it holds the potential to provide highly actionable prognostic information that may be integrated into an institution’s practices for improving retention. If a careful account is made of the particular retention interventions actually experienced by individual students then comparisons of estimated risk versus actual outcomes can be used to evaluate specific retention interventions. Retention practices themselves may thus become working hypotheses, constantly being tested and modified over time in a process of adaptive management and continuous improvement. Furthermore, the collected data may be incorporated into survival analyses conducted in subsequent years in order to account for retention practice efficacy on an ongoing basis. Finally, given that risk estimates are expressed as probabilities, with values that always exist in the range of 0 to 1 and have consistent interpretations, it ought to be possible to directly compare risk assessments across analyses and thus across institutions. In its fully realized form, then, the approach advocated in this report integrates retention practice design with assessment and also integrates retention practice assessment with enrolment forecasting.

It seems, though, that to actually achieve the potential benefits of using the approach, one must start with a fairly sophisticated, student-level data set. The current gap between available data and what is required to produce precise retention risk estimates is an issue to be grappled with. The advantages of the approach are many, but some qualitative research and much data collection work must be accomplished in order to make the approach fully viable.

Appendix A: Pilot Study Details

Data

Time-to-event data were compiled for all newly enrolling, first-year, direct-entry undergraduate degree students entering York University over the study period running from fall 1996 through fall 2006. The data set includes one record for each of the 83,593 students conforming to these criteria who enrolled within this timeframe. Only students who were fully registered as of the official autumn enrolment count date (November 1) are considered enrolled for a particular academic year.

The data set includes two data elements that together form a “time-to-event couplet.” The first element of the couplet is the number of elapsed academic years from the time of entry to the first occurrence of the event of interest, which in this case is a stop-out. In this pilot study, the term “stop-out” is defined as a failure to re-enroll at the university (as of the official November 1 count date) either temporarily or permanently. Thus, stop-out is an inclusive term encompassing temporary and permanent leavers from York University without regard for their eventual status, if any, at other PSE institutions. In the case of students who stop-out, return and subsequently stop-out again, only the time to the first stop-out event is considered. A student who stops-out is assumed to have remained enrolled for the duration of the entire academic year prior to the stop-out. Thus, a student who entered in fall 2000 but did not return the next year – and was thus observed as not enrolled on November 1, 2001 – would generate a time-to-event value of 1 academic year. (This procedure is due to the temporal resolution of the data that was available for use in the pilot study and is certainly not a requirement of the analysis approach in general. In reality, students may stop-out at any point during an academic year; stop-out-generating processes are temporally continuous, not discrete.)

Censored observations are included in the data set, and censoring occurs either because students were still enrolled at the end of the study period or because they “withdrew from the study” earlier on by virtue of having graduated. (More will be said about this second source of censored data in the next section.) In the case of censored data, the time-to-event element provides the number of academic years over which the student was “under observation.” Students taking four years to complete their degree, for example, would generate a time-to-censoring value of 4 academic years. On the other hand, the status of students still enrolled at the end of the study period are, by definition, not observed at the beginning of the academic year immediately after the end of the study period, and thus an almost – but not fully – complete increment is added to the students’ time-to-censoring values for their last year under observation. For example, students who were enrolled for the last three years of the study period receive times-to-censoring values of 2.9 academic years, not 3. A full year increment is added only if a student’s status is observed on November 1 of the subsequent year, and such an observation is not possible past the end of the study period. The actual proportion of the final year that is “shaved off” is inconsequential, given the fact that the Nelson-Aalen estimator of the cumulative hazard function is used in the analysis. It only matters that students still enrolled up to the end of the study period receive a partial, not full, increment for the final year under observation. The second element of the time-to-event couplet is a flag indicating whether the observation is censored or whether it is an actual stop-out event.

The time-to-event couplet is the response variable. The remaining elements of the data set are predictor variables taken from the admissions, biographical, financial and academic data that York University maintains for each student. A number of the predictor variables are represented by two data elements: one for the predictor's value during a student's initial year of study and a second element for the predictor's value during a student's most recent academic year, the one immediately prior to the stop-out or censoring event. Table 2 provides an enumeration of data elements eventually included as predictor variables in the statistical models reported in this report.

Table 2
List of data elements used as predictor variables in the statistical models reported in this paper.

Initial year: OAC grade group (60-69.9%, 70-74.9%, 75-80.9%, 80-89.9%, 90-100%)
Initial year: full-time status (at least 80% of full load)
Initial year: full-time status (at least 60% of full load)
Initial year: age (years)
Initial year: living in residence/commuting status
Initial year: domestic/international student status
Initial year: admissions adjudication type (101 or 105)
Current year: previous stop-out flag
Current year: academic year level repeater
Current year: domestic/international student status
Current year: domestic/international status switch flag (0 if no switch since initial year; otherwise 1)
Current year: living in residence/commuting status
Current year: honours/general curriculum status
Current year: cumulative 1st major count (initially 1, increments with every switch)
Current year: cumulative 2nd major count (initially 1, increments with every switch)
Current year: home faculty
Current year: faculty switch flag (0 if no switch since initial year; otherwise, 1)
Current year: degree objective (BA, BSc, etc.)
Current year: degree objective switch flag (0 if no switch since initial year; otherwise, 1)
Current year: full-time status (at least 80% of full load)
Current year: full-time status (at least 60% of full load)
Current year: fiscal full-time equivalent value
Current year: 1st major
Current year: bursary count
Current year: bursary amount
Current year: scholarship count
Current year: scholarship amount

Of the 83,593 students included in the data set, 26 percent stayed enrolled until the end of the study period, 36 percent stayed enrolled until graduation and 38 percent stopped-out (see Figure 4). Furthermore, it is known that 20 percent of the 31,850 students who stopped-out eventually returned to York University, and 35 percent of those who returned eventually graduated. It should be noted that the data set includes students who entered York University as part-time students, which generates proportionately more stop-outs than would a full-time-only data set. Amongst the students in this pilot study, only 34 percent of those entering York as full-time students (with at least 80 percent of full course load) eventually stopped-out, compared to 60 percent of those entering as part-time students. In the case of students who entered between fall 1996 and fall 2000, 40 percent of all initial stop-outs occurred during the first year of studies, and a total of 65 percent occurred by the end of second year (see Figure 5). Students who stop-out during first year are less likely to return to York University than are students whose initial stop-out occurs later on in their studies: only 21 percent of first year stop-outs subsequently returned to York, whereas 35, 37 and 28 percent of initial stop-outs occurring in second, third and fourth years, respectively, returned to their studies at York University.

Figure 4
Enrolment pathway diagram for the students included in the retention risk analysis pilot study data set.

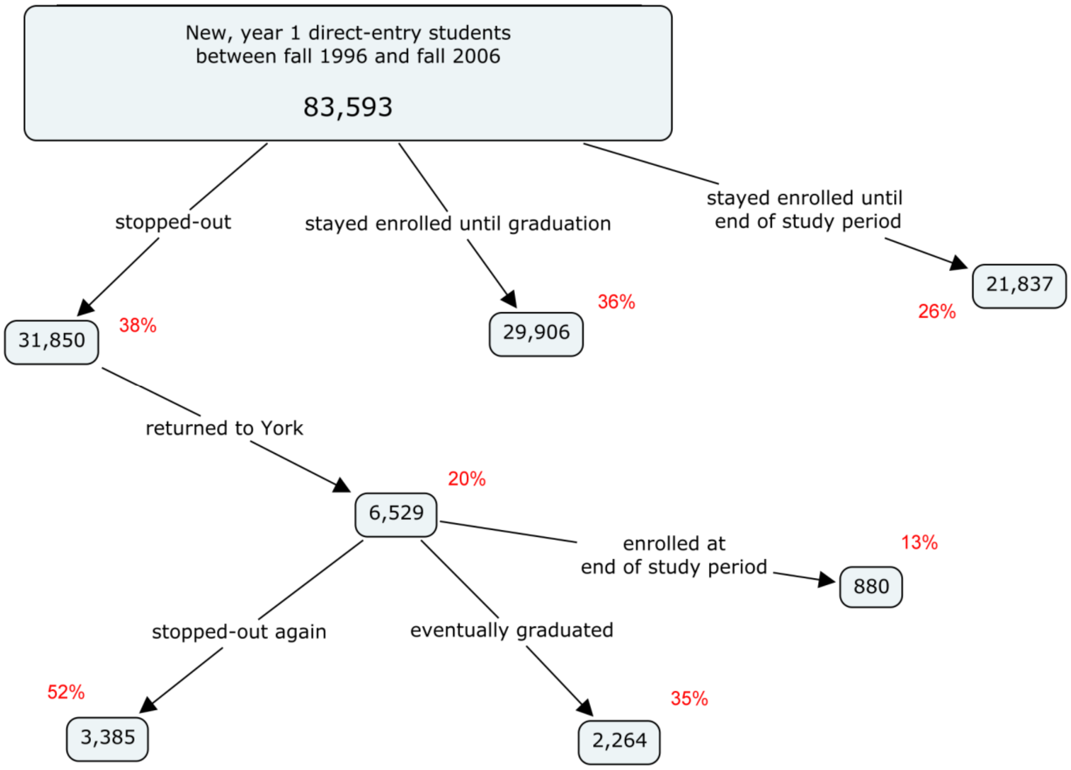
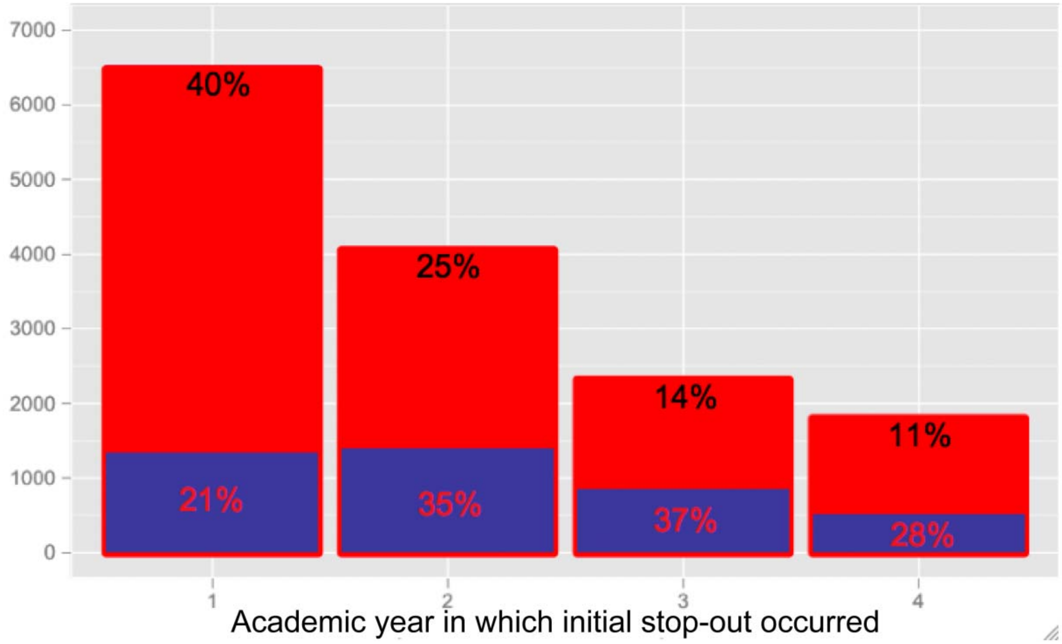


Figure 5
Histogram of initial stop-outs by the academic year in which it occurred for the first 4 years from the time of first entry into undergraduate studies. Data is for students in the pilot study data set who entered between fall 1996 and fall 2000. Red bars represent the number of initial stop-outs that occurred during a particular year (with percentage of all stop-outs – including those occurring beyond the first 4 years – shown in black type) and blue bars represent the number of stop-outs occurring during a particular year who eventually re-enrolled at York University (with percentage re-enrolling shown in red type).



Analytical Methods

The time-to-event data set was analyzed using the “random survival forests” (RSF) technique, which is an ensemble survival tree method for the analysis of right-censored time-to-event data (Ishwaran and Kogalur, 2007; Ishwaran, Kogalur, Blackstone, & Lauer, 2008). Survival tree methods are an extension of classification and regression tree methods to survival data (i.e., time-to-event data). Ensemble tree methods depend on combining in a predetermined way the fitted values from a large number of tree-growing attempts. A stochastic algorithm produces a large ensemble of trees (hence “forest”) from which output is combined. The idea is that a weak procedure (e.g., one that produces individual survival trees exhibiting poor prediction performance) becomes strong when operating “by committee,” and, indeed, ensemble methods perform extremely well in many cases (Berk, 2006; Breiman, 2001). The RSF algorithm used in this analysis is summarized as follows, after Ishwaran, Kogalur, Blackstone, & Lauer (2008):

1) Draw many bootstrap samples from the full, original data set. Each bootstrap sample excludes about 37 percent of the records from the original data. (The excluded data for any one sample are referred to as being “out-of-bag” for that sample.)
2) Grow a binary survival tree for each bootstrap sample. At each node of the tree, randomly select several candidate predictor variables. The node is split using the candidate variable that maximizes survival differences between daughter nodes, based on a predetermined survival criterion. By maximizing survival difference, the tree pushes dissimilar cases apart. As the tree grows and dissimilar cases are separated, each node becomes more homogeneous in terms of survival outcomes.
3) Grow each tree to full size under the constraint that the terminal nodes should have no less than a predetermined number of unique, non-censored events.
4) Calculate a cumulative hazard function (CHF), using the Nelson-Aalen estimator, for each terminal node in every tree.
5) Calculate an ensemble CHF for each record by “dropping” the records down each tree. Because the trees are binary, each record is guaranteed to fall into a single terminal node per tree. The CHF associated with that node is assigned to the record. The ensemble CHF for the record is the average of all CHFs assigned to that record across all trees.
6) The ensemble CHF predictions are biased because the same data used to generate predictions were used to grow the trees in the first place. Generate nearly unbiased CHF predictions by dropping records down only those trees for which the records were out-of-bag and therefore not involved in the growing process. These nearly unbiased estimates are called out-of-bag (or OOB) ensemble CHFs.
7) New prediction data can be dropped down the trees in order to predict ensemble CHFs for the new data.

This technique was selected because it is easy to implement in a prospective forecast setting, its prediction performance compares favourably with other methods (particularly when there are many predictor variables) and it can handle interaction effects, nonlinearities in the relationship between predictor and response variables and collinearities amongst predictors. The main output of the technique, when used in a forecast setting with new data, is a set of predicted ensemble cumulative hazard functions (CHFs), one for each record in the new data set. The CHFs are easily converted to survival probability functions, using well-understood relations between the various quantities used to describe survival data (Klein and Moeschberger, 2003). In particular, the exponentiated negative of a cumulative hazard value is taken as its equivalent ensemble survival probability value. In the context of this pilot study, these survival probabilities can be interpreted as student-specific retention risk estimates. The complements of the survival probabilities are interpretable as student-specific predictions of stop-out risk.

For the pilot study, five separate “training set” samples of 10,000 records were drawn from the entire data set of 83,593 records. Each training set was drawn without replacement from the original full data set, such that data for a particular student might be included in more than one training set but would not occur more than once within a single training set. The RSF algorithm was then applied to each of the five training sets and output compared. Output from one of the five training sets was then selected for further analysis: out-of-bag CHF estimates for the selected training set were converted to out-of-bag survival probability estimates, and these, along with actual, known time-to-event values, were input to a binary classification tree algorithm in order to quickly identify threshold survival probability values below which students might be classified as being “at risk” of stopping-out. A “test set” of 10,000 records, none of which occurred in the selected training set, was then drawn from the original 83,593 records. Ensemble survival probability estimates were generated for each record in the test set by dropping the records down the ensemble of trees grown using the training data, and these estimates were used in conjunction with the previously obtained threshold values to classify students as “stop-outs” or “continuers.” Prediction performance of the classifier was assessed by comparing classifications against actual, known stop-out status for students represented in the test set. (In a fully operationalized study, the RSF tree ensembles and survival threshold values would be applied to entirely new data, for which times-to-event had not yet been observed, in order to generate prospective predictions.)

The accuracy of the ensemble survival probability estimates that were generated for the test data was assessed by comparing the probability estimates with actual, “localized” survival rates. Actual, local rates were calculated using a moving window algorithm: to find the actual rate associated with a particular survival probability value, all observations in the test data set were first weighted according to their Euclidean distance from that value (which we will call the target value). Observations with estimated survival probabilities that are more than ±0.1 distant from the target value were given a weight of zero. The remaining, less distant, observations were given a weight equal to the complement of their distance from the target value raised to the 4th power (to emphasize observations very close to the target value). For example, given a target survival probability of 0.5, all observations with estimated survival probabilities greater than 0.6 or less than 0.4 would receive a weighting of zero. An observation with a survival probability estimate of 0.525 is 0.025 distant from the target value of 0.5. The complement of this distance is 0.975, and the weight assigned to the observation would be $0.975^4 = 0.9037$. The weighted average of a binary variable indicating those who continue past first year was then calculated and interpreted as the actual localized survival rate associated with the target survival probability. One such rate was calculated for each ensemble survival probability estimate value generated via the test data, and a plot of probability estimates versus actual localized rates was produced.

An add-on package for R statistical software (R Development Core Team, 2008), called “randomSurvivalForest 3.5.1” (Ishwaran and Kogalur, 2007), is the reference implementation of the RSF technique and was used in this pilot study. Settings required by randomSurvivalForest were set as listed in Table 3. Although the software does offer facilities for imputing missing data, the decision was made to merely exclude records with missing data from the pilot study analysis. The R package “rpart” (Therneau, Atkinson, & R port by Brian Ripley, 2009) was used for the binary classifications.

As previously mentioned, in this study, censored data arises in one of two ways. Either a student was still enrolled at the end of the time period over which data were collected or the student “withdrew” from the study by virtue of having graduated. Treating the latter category of students as censored observations makes sense if one considers the graduation event as being outside the scope of interest of the study and independent of the process that leads to stop-outs (i.e., the time-to-graduation should not convey any information about how long the student would have remained enrolled had the student not graduated). This is akin to assuming a university system in which there is no such thing as graduating and students take courses until they decide to stop-out (Klein and Moeschberger, 2003). Let this particular formulation of the time-to-event modelling problem be called the “single risk formulation.” Strictly speaking, this is the formulation of the problem that is used in this pilot study.

On the other hand, the analysis in this pilot study is nearly equivalent to a “competing risks formulation,” in which multiple types of events are acknowledged – in this case, either an initial stop-out or a graduation – but only one of the types (namely, the stop-out type) is treated as interesting given the research question at hand, and the timeframe of interest when interpreting results is restricted to the first two academic years. The timeframe restriction is practical, since, as noted previously, most stop-outs occur prior to the beginning of the third academic year, and, from a retention management point of view, early interventions (prior to second year) are usually attempted. Since one of the two event types – graduation – almost never occurs prior to the end of the third academic year, a competing risks formulation produces results that are nearly equivalent to those of the single risk formulation, given the timeframe restriction supplied above.

Indeed, two quantities often used in a competing risks formulation – the complement of the cause-specific survival function and the cause-specific cumulative incidence function – will equal each other for a particular event type when they are calculated for time points before which any events of the other competing types have occurred (Klein and Moeschberger, 2003). In the current context, the stop-out-specific quantities will equal each other when calculated for time points before which any graduations have occurred. And since the cause-specific survival function for a particular event type is computationally identical to the survival function in a single risk formulation focusing on that same event type, it is clear that the single risk formulation will provide results that are very close to a competing risks formulation, again given the specified timeframe restriction.

In any event, adopting the single risk formulation for the pilot study was a matter of necessity; software implementing an ensemble survival tree method for competing risk formulations was not yet available at the time of analysis, although the most recent major upgrade to the RSF algorithm includes the new ability to fully analyze competing risks data (Ishwaran and Kogalur, 2010). While some results for time points beyond the second academic year are provided, emphasis will be placed on results pertaining to stop-outs occurring prior to the third year, before the vast majority of graduations take place.

Results

The RSF algorithm was run once on each of the five training sets, and summary information regarding these runs is provided in Table 3. During each run, prediction error rates stabilized after about 400 survival trees had been grown, and each run was stopped after the growth of 600 survival trees. An overall measure of prediction performance for a run of the RSF algorithm is obtained using Harrell’s concordance index (C-index), the algorithmic details of which are described by Harrell, Califf, Pryor, Lee, & Rosati (1982). Details regarding the use of the C-index in RSF are provided by Ishwaran, Kogalur, Blackstone, & Lauer (2008). C-index values range from 0 to 1, with values of 0.5 indicating prediction performance that is no better than guessing and values of 0.0 indicating error-free predictions.

Table 3
Summary information regarding each run of the random survival forest algorithm. Data for each run consisted of 10,000 randomly selected records (without replacement) taken from the full data set of 83,593 time-to-stop-out records.

	Run 1	Run 2	Run 3	Run 4	Run 5
Number of bootstrap samples (trees per forest)	600 bootstrap samples				
Total number of predictor variables	27 variables				
Number of predictors randomly selected at each node split	5 variables				
Splitting rule	logrank random				
Number of random split points	1 split point				
Minimum terminal node size	3 records				
Sample size after removing records with missing data	9,925	9,918	9,923	9,930	9,906
Number of actual stop-outs	3,835	3,762	3,845	3,809	3,716
Average number of terminal nodes	219	204	210	217	209
Estimated prediction error rate (C-index)	19.2%	18.2%	20.1%	19.1%	19.2%

The C-index values for all five runs of the RSF algorithm, expressed as percentages (Table 3), range between 18.2 and 20.1 percent. C-index values in this range are certainly better than guessing, but at the same time, are not exceptionally low. It is important to note, however, that the C-index values indicate overall prediction performance for stop-out events occurring across all time points and do not provide information about predictions based on a specific time point. In practice, one would be interested in probabilities associated with stopping-out after 1 academic year or perhaps 2 years. It is therefore relevant to evaluate the prediction performance of the RSF algorithm when addressing these more specific questions.

Figure 6 gives results from the test set analysis, illustrating the relationship between ensemble estimates of the probability of surviving (i.e., continuing with studies) beyond the first academic year and the actual occurrence of stop-outs up to the end of first year. It is clear from the figure that the survival probability estimates provide good discrimination between those who stop-out by the end of first year and those who continue with their studies. Indeed, a threshold ensemble survival probability of 0.72 provides good prediction performance when applied to the test set. A confusion matrix (Table 4) showing details of the prediction performance indicates an overall prediction error of 9.2 percent, a prediction sensitivity of 85.3 percent (i.e., 14.7 percent of actual first-year stop-outs incorrectly classified as continuing students) and a prediction specificity of 69.6 percent (30.4 percent of predicted first-year stop-outs actually continuing uninterrupted into second year). Only 8.0 percent of students actually continuing past first year were incorrectly classified as first-year stop-outs.

Figure 6

Ensemble estimates of the probability of surviving beyond first year, by first year stop-out status. Red points indicate stop-outs regardless of the year in which they occur, and blue points indicate non-stop-outs. The red line represents a survival probability threshold-value of 0.72 below which students are predicted as first year stop-outs. Observations censored due to students still being enrolled in first year at the end of the study period are excluded from the confusion matrix, since the actual status for these first-year students at the beginning of the subsequent academic year (i.e., the year immediately following the end of the study period) are unknown. Random noise has been added to the location of points along the x-axis to reduce over-plotting.

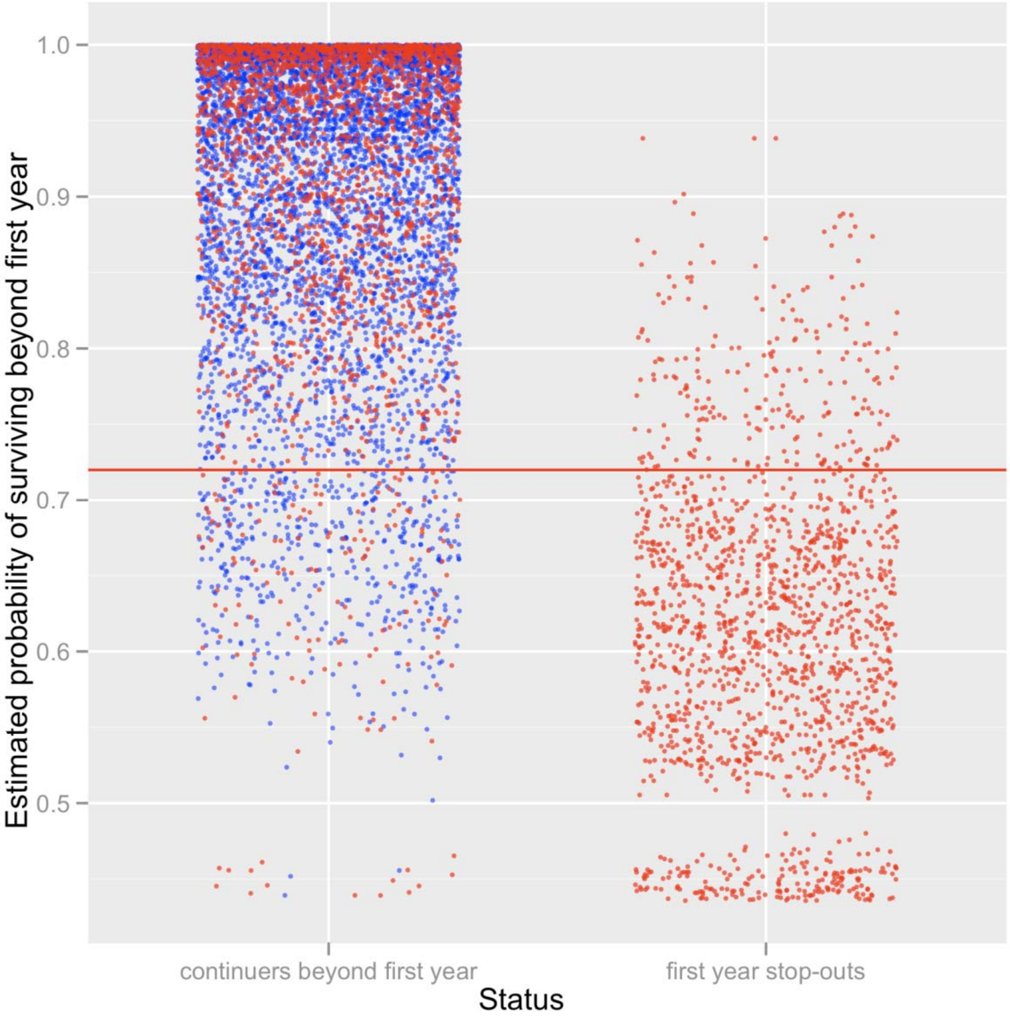


Table 4
Confusion matrix for a classification of students into two groups: “first-year stop-outs” and “continuers.” The classification is based on a threshold value of 0.72 applied to ensemble estimates of the probability of surviving beyond first year, for students represented in the test data set. Observations censored due to students still being enrolled in first year at the end of the study period are excluded from the confusion matrix, since the actual status for these first-year students at the beginning of the subsequent academic year (i.e., the year immediately following the end of the study period) are unknown.

		Predicted		
		Continuers	First-year stop-outs	Total
Observed	Continuers	6,911	601	7,512
	First-year stop-outs	238	1,376	1,614
	Total	7,149	1,977	9,126

The left column of points in Figure 6 includes a large number of upper-year stop-outs (coloured red), only a small portion of which is entrained as “by-catch” in the classification of first-year stop-outs (i.e., reside below the 0.72 first-year survival probability threshold). Since about 25 percent of all initial stop-outs occur during second year and 60 percent occur during one of the upper years (including second year), an assessment of prediction performance for second-year and upper-year stop-out classifications is also of interest. To accomplish this assessment, a classification was produced that predicts stop-outs – regardless of the year in which the stop-outs occur – based on estimated probabilities of survival beyond two years. At first glance, the results for this classification (see Figure 7 and Table 5) are not particularly impressive at first glance, with an overall prediction error of 30.4 percent, and a prediction sensitivity of 52.5 percent (i.e., 47.5 percent of actual stop-outs incorrectly classified as non-stop-outs). On the other hand, prediction specificity is a respectable 82.4 percent (i.e., only 17.6 percent of predicted stop-outs actually continue uninterrupted to graduation) and 12 percent of students actually continuing their studies uninterrupted until to graduation were incorrectly classified as stop-outs.

What makes the results of this classification attempt more appealing, however, is that fully 1,458, or 90.3 percent, of first-year stop-outs in the test dataset (represented by red dots in Figure 7) and 363, or 35.3 percent, of second-year stop-outs (represented by green dots) are correctly predicted. These results, combined with the relatively high overall prediction specificity of 82.4 percent, suggest that it should be possible to develop stop-out prediction models that are powerful enough to deploy in an operational setting for the purposes of triaging student retention risk (particularly amongst first year students) and directing retention interventions accordingly.

Figure 7
Ensemble estimates of the probability of surviving beyond 2 academic years, by stop-out status. Red points indicate stop-out events occurring during first year, green points indicate stop-out events occurring during second year, yellow points indicate stop-outs occurring at some point after second year and blue points indicate students who continue uninterrupted to graduation. The red line represents a survival probability threshold-value of 0.67, below which students are predicted as stop-outs. Observations censored due to students still being enrolled (and not graduating) at the end of the study period are excluded from the figure, since the actual status for these students at the beginning of fall 2007 (i.e., the year immediately following the end of the study period) are unknown. Random noise has been added to the location of points along the x-axis to reduce over-plotting.

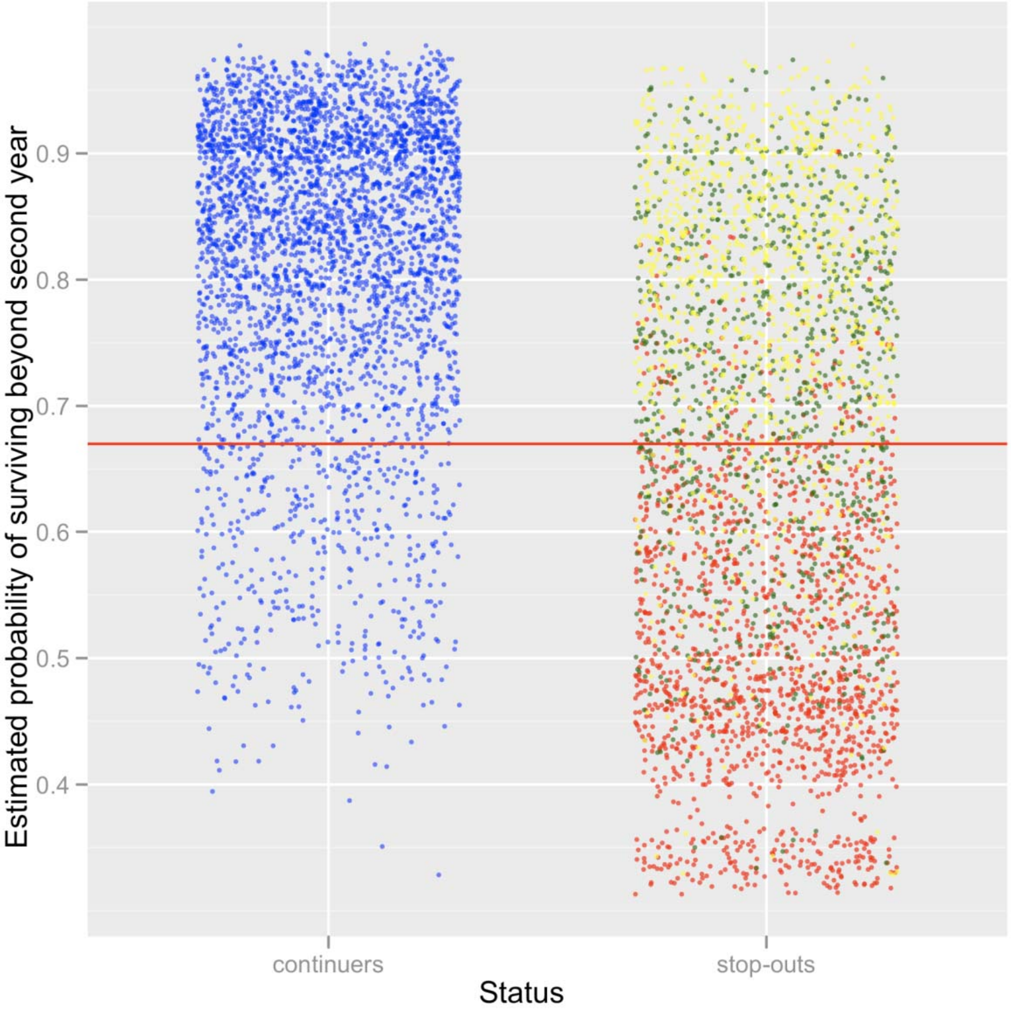


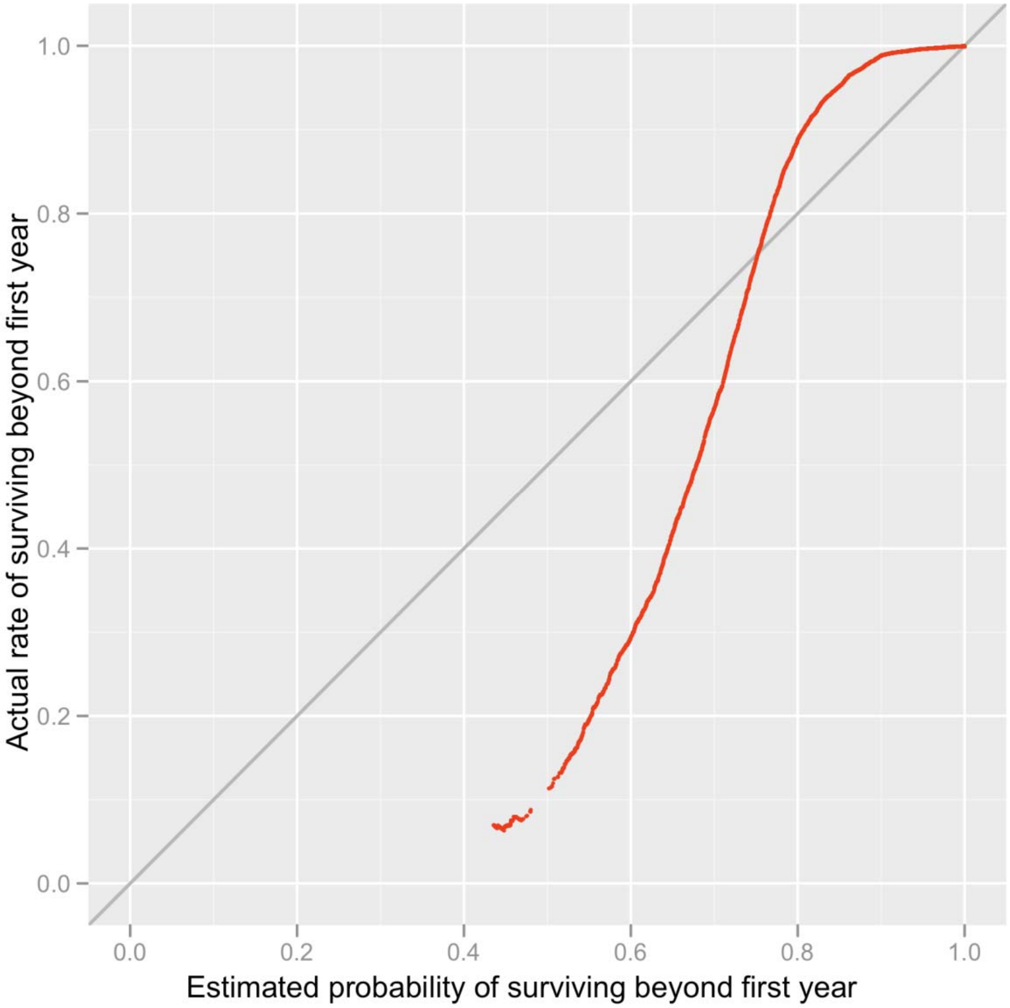
Table 5
Confusion matrix for a classification of students into two groups: “stop-outs” and “continuers.” The classification is based on a threshold value of 0.67 applied to ensemble estimates of the probability of surviving beyond 2 years, for students included in the test data set. Observations censored due to students still being enrolled (and not graduating) at the end of the study period are excluded from the confusion matrix, since the actual status for these students at the beginning of fall 2007 (i.e., the year immediately following the end of the study period) are unknown.

		Predicted		
		Continuers	Stop-outs	Total
Observed	Continuers	3,127	428	3,555
	Stop-outs	1,808	1,998	3,806
	Total	4,935	2,426	7,361

Figure 8 illustrates the relationship between estimated probabilities of survival past the first year of studies and actual rates of survival past the first year. In a general sense, the survival probability estimates are well behaved: higher probability values are associated with higher actual rates, allowing for effective binary classifications of students into “continuers” and “stop-outs.” On the other hand, the plot clearly shows that the estimated survival probabilities generated by the RSF algorithm do not closely match actual rates except in two small regions of the plot (one of these regions is centered on the value 0.75 and the other near the value 1.0). In particular, students most at risk of stopping-out have estimated survival probabilities that are much higher than the actual localized survival rates. This situation reduces the usefulness of the estimated probabilities for the purposes of assessing institutional retention risk profiles and for evaluating the efficacy of retention practices, although it is reasonable to suspect that developing and using more sophisticated data sets than the one used in this pilot study would yield more accurate survival probability (i.e. retention risk) estimates.

Figure 8
Plot of actual, localized rates of surviving beyond the first year of studies, by estimated probabilities of surviving beyond the first year of studies. Actual, localized rates are calculated using a moving window algorithm: To find the actual rate associated with a target survival probability value, all observations were first weighted according to their distance from the target value. Observations with estimated survival probabilities that are more than ±0.1 distant from the target value were given a weight of zero. The remaining, less distant observations were given a weight equal to the complement of their distance from the target value raised to the 4th power. A weighted average of a binary variable indicating survival past first year was then calculated and interpreted as the actual, localized survival rate associated with the target survival probability. Actual rates were calculated for every survival probability value generated

for the test data set. The black line in the plot references equality between actual rates and estimated probabilities.



References

Astin, A.W. (1993a). *What matters in college?: Four critical years revisited*. San Francisco: Jossey-Bass Publishers.

———. (1993b). College retention rates are often misleading. *Chronicle of Higher Education, 22, A48.*

Astin, A.W. (1997). How “good” is your institution's retention rate? *Research in Higher Education, 38,* 647-658.

Astin, A., & Oseguera, L. (2005). *Degree attainment rates at American colleges and universities: Effects of race, gender, and institutional type* (rev. ed.). Los Angeles: Higher Education Research Institute.

Auclair, R., Bélanger, P., Doray, P., Gallien, P., Groleau, A., Mason, L, & Mercier, P. (2008). *Transitions* **–** *First generation students: A promising concept?* Research paper, Canadian Millenium Scholarship Foundation, Montreal.

Berk, R.A. (2006). An introduction to ensemble methods for data analysis. *Sociological Methods and Research, 34,* 263-295.

Bourdieu, P. (1977). *Outline of a theory of practice*. Cambridge and New York: Cambridge University Press.

———. (1990). *The logic of practice*. Stanford: Stanford University Press.

Breiman, L. (2001). Random forests. *Machine Learning, 45,* 5-32.

Cabrera, A. F., Nora, A., & Castaneda, M.B. (1993). College persistence: Structural equations modeling test of an integrated model of student retention. *Journal of Higher Education, 64,* 123-139.

Canadian Career Development Foundation. (2007). *Applying the construct of resilience to career development: Lessons in curriculum development*. Montreal: Canadian Millenium Scholarship Foundation.

DesJardins, S. L., McCall, B.P., Ahlburg, D.A., & Moye, M.J. (2002). Adding a timing light to the "tool box." *Research in Higher Education, 43,* 83-114.

Duggan, M.B. (2002). The effect of social capital on the first-year persistence of first generation college students. Ed.D. dissertation, Boston: University of Massachusetts.

Finnie, R., & Qui, H.T. (2008). *The patterns of persistence in post-secondary education in Canada: Evidence from the YITS-B dataset*. MESA Project Working Paper, Educational Policy Institute, Queen's University, Kingston.

Grayson, J.P., & Grayson, K. (2003). *Research on retention and attrition*. Montreal: Canadian Millenium Scholarship Foundation.

Harrell, F. E., Jr., Califf, R.M., Pryor, D.B., Lee, K.L., & Rosati, R.A. (1982). Evaluating the yield of medical tests. *Journal of the American Medical Association, 247,* 2543-2546.
Higher Education Research Institute. (2003). How "good" is your retention rate? Using the CIRP survey to evaluate undergraduate persistence. Los Angeles: Higher Education Research Institute.

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

Ishitani, T.T., & DesJardins, S.L. (2003). A longitudinal investigation of dropout from college in the United States. *Journal of College Student Retention, 4,* 173-201.

Ishitani, T.T., & Snider, K.G. (2006). Longitudinal effects of collegepreparation programs on college retention. *IR Applications, 9*, 1-9.

Ishwaran, H., & Kogalur, U.B. (2007). Random survival forests for R. *R News, 7,* 25-31.
———. (2010). *randomSurvivalForest: Ishwaran and Kogalur's Random Survival Forest.* http://CRAN.R-project.org/package=randomSurvivalForest.

Ishwaran, H., Kogalur, U.B., Blackstone, E.H., & Lauer, M.S. (2008). Random survival forests. *The Annals of Applied Statistics, 2,* 841-860.

Klein, J.P., & Moeschberger, M.L.(2003). Nonparametric estimation of basic quantities for right-censored and left-truncated data. In *Survival analysis: Techniques for censored and truncated data* (2nd ed., pp. 91-138). New York: Springer.

Lehmann, W. (2007). "I just didn't feel like I fit in": The role of habitus in university dropout decisions. *The Canadian Journal of Higher Education, 37,* 89-110.

Lohfink, M.M., & Paulsen, M.B. (2005). Comparing the determinants of persistence for first-generation and continuing-generation students. *Journal of College Student Development, 46,* 409-428.

London, H.B. (1996). How college affects first-generation students. *About Campus* 1, no. 5: 9-13.

Parkin, A., & Baldwin, N. (2009). *Persistence in post-secondary education: The latest research*. Research note. Montreal: Canadian Millennium Scholarship Foundation.

Pascarella, E.T., Pierson, C.T., Wolniak, G.C., & Terenzini. P.T. (2004). First-generation college students: Additional evidence on college experiences and outcomes. *Journal of Higher Education, 75,* 249-285.

R Development Core Team. (2008). *R: A language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing.

Radcliffe, P.M., Huesman, R.L., Jr., Kellogg, J.P., & Jones-White, D.R. (2009). Identifying students at risk: Utilizing survival analysis to study student-athlete attrition. *IR Applications, 21,* 1-15*.*

Sandler, M.E. (2000). Career decision-making self-efficacy, perceived stress, and an integrated model of student persistence: A structural model of finances, attitudes, behavior, and career development. *Research in Higher Education, 41,* 537-580.

Therneau, T.M., Atkinson, B., & R port by Brian Ripley. (2009). Rpart: Recursive partitioning and regression trees. Vienna: R Foundation for Statistical Computing.

Thomas, S.L. (2000). Ties that bind: A social network approach to understanding student integration and persistence. *Journal of Higher Education, 71,* 591-615.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research, 45,* 89-125.

———. (1993). *Leaving college: Rethinking the causes and cures of student attrition* (2nd ed.). Chicago: University of Chicago Press.

York University. (2007). 2006-07 multi-year accountability agreement report-back: York University. Toronto: York University.