International Journal of Engineering Education Vol. 29, No. 1, pp. 132–138, 2013
Printed in Great Britain
0949-149X/91 $3.00+0.00

Modelling Engineering Student Academic Performance Using Academic Analytics*

STUART PALMER
Faculty of Science and Technology, Deakin University, Geelong, Victoria, Australia 3220. E-mail: spalm@deakin.edu.au

Internationally, the recruitment, management and retention of students has become a high priority for universities. The use of information technology systems and student data by institutions to understand and improve student academic performance is often referred to as 'academic analytics'. This paper presents an academic analytics investigation into the modelling of academic performance of engineering students enrolled in a second-year class. The modelling method used was binary logistic regression, and the target predicted variable was 'success status'—defined as those students from the total originally enrolled group that achieved a final unit grade of pass or better. This paper shows that student data stored in institutional systems can be used to predict student academic performance with reasonable accuracy, and it provides one methodology for achieving this. Importantly, significant predictor variables are identified that offer the ability to develop targeted interventions to improve student success and retention outcomes.

Keywords: academic analytics; student academic performance; engineering education; binary logistic regression

1. Introduction

Internationally, the recruitment, management and retention of students has become a high priority for universities [1–2]. These issues have been noted as acute and longstanding for engineering education [3–4], and many reasons have been posited for the observed difficulties in attracting and then retaining engineering students through to successful graduation [5]. Once students are enrolled, understanding and quantifying student retention and persistence is not necessarily straightforward. While there is significant published research on the topic of student academic performance, it has been noted that a wide range of definitions exist for the terminology in this area [6]. In the case of engineering education, an examination of a large, multi-institution student enrolment data set revealed that while up to 40 percent of students enrolling in engineering leave the course in the first year of their study, many of these students intentionally take up another course of study at the same institution, or complete an engineering qualification at another institution [7]. In the same study, the percentages of students still enrolled in an engineering program into their eighth semester after entry varied widely between the participating institutions. Another large, multi-institution investigation found that the completion/graduation rates of commencing engineering students varied dramatically depending on the number of years since enrolling in engineering, only approaching a stable final value at six years after original enrolment [8]. A conclusion from the literature is that there is a wide range of terminology associated with student academic performance (retention, progression, persistence, wastage, completion, etc.), that there is a wide range of possible measures of student academic performance, and that it is important to be clear about the particular measure(s) being employed.

Significant research has been conducted into the factors contributing to, and predictors of, student academic performance over a long period of time [6, 9]. Historically, much of this research has been based on qualitative surveying of students cohorts as they progress, or not, through their studies. However, while this work has been valuable for formulating and validating theories of student academic performance, the practical utility of survey-based approaches has been questioned on the grounds of lack of generalizability of results and the costliness of conducting such surveys [1]. A key characteristic of a useful model for understanding student academic performance is that it moves beyond abstract theoretical concepts and translates into practical actions [6]. Although relatively new to higher education, many other sectors have been employing data mining techniques for many years to understand the factors that assist them to retain 'customers', as the cost of keeping existing customers is generally much lower than the cost of recruiting new ones [1]. Similar cost advantages in retaining existing students have been observed in higher education [3]. The use of information technology (IT) systems and student data by institutions to understand and improve student academic performance is often referred to as 'academic analytics' [10]. An investigation to compare a traditional survey-based retention research methodology with a data mining/analytics approach that used existing

* Accepted 4 October 2012.

student data held by the university found that the data-driven approach outperformed the survey approach in predictive utility, and also produced a much simpler model [11]. Much of the research relating to analytics-style approaches to prediction of student academic performance has been based on the use of general student information contained in institutional databases. However the almost ubiquitous presence of learning management systems (LMSs) in higher education [12], and the vast amounts of data on student engagement with learning resources and activities that they hold means that they are emerging as a key source of data for the prediction of student academic performance [13–14].

Many approaches to the predictive modelling of student academic performance can be found in the literature, including: linear regression [13]; logistic regression [8, 11, 13]; structural equation modelling [15–16]; and data mining/machine learning techniques [1–2, 14]. Logistic regression is a relatively simple procedure supported by many statistical and numerical analysis systems, and is preferred over linear regression in higher education applications where the independent predictor variables may be either continuous/interval or discrete/categorical, and the dependent predicted variable is binary categorical, i.e., pass/fail, retained/lost, etc. [8, 11]. In logistic regression, the predicted value $f(z)$ of a predictor input is modelled by a logistic function of the form given in Eq. 1.

$$f(z) = \frac{e^z}{e^z + 1} = \frac{1}{1 + e^{-z}} \tag{1}$$

The predictor input is formed from a function of the form given in Eq. 2.

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \tag{2}$$

Here, $x_1 \ldots x_k$ are the set of predictor variables; and $\beta_0 \ldots \beta_k$ are the model coefficients determined by the logistic regression algorithm. The logistic function has the form of a sigmoid curve that has asymptotes of 1 at $z(\infty)$ and 0 at $z(-\infty)$, is linear in the region z = 0.5, and approximates $f(z)$ as a binary variable.

In investigations of the prediction of student success/retention many factors are identified as significant predictors, but one type of measure stands out for its repeated identification—that is, measures of student prior academic performance, typically grade point average (GPA) or some similar measure [1–2, 8–9, 15, 17]. GPA is an example of a time invariant predictor variable; though it may vary in the longer term. While GPA has been shown many times over to be a significant a priori predictor of future student academic performance, it has also been shown that the overall predictive power of models of student academic performance can be significantly enhanced by the inclusion of time varying student data relating to students' current study activities [14]. This may be due to there being an upper limit on how much variation in student performance can be predicted from pre-existing factors, and that the on-going choices that students make also significantly impact on their academic performance [8]. The almost ubiquitous presence of LMSs in higher education suggests them as an obvious source of real-time data relating to student study activity, and the inclusion of LMS usage data has been shown to improve the performance of models of student academic performance [13]. A large majority of the published research on retention and persistence relates to students in their first ('freshman') year of university life—and while this is a critical transition period in a student's academic career, there is a need to consider the predictors of student academic performance in their second year and beyond [2].

This paper presents an academic analytics investigation into the modelling of academic performance of engineering students enrolled in a second-year class. The modelling method used was binary logistic regression, and the target predicted variable was 'success status'—defined as those students from the total originally enrolled group that achieved a final unit grade of pass or better. The modelling exercise draws on both time invariant demographic data from student information systems, and time varying data from the institutional LMS.

The investigation seeks to establish the feasibility of such modelling, and to identify key predictive variables that could be used to target practical and timely interventions that could improve overall student academic performance outcomes.

2. Context

The School of Engineering at Deakin University in Australia offers undergraduate and postgraduate engineering programs. These programs are delivered in both on-campus and off-campus modes. Off-campus students are typically mature aged, working full-time, have significant experience in an engineering-related job role, and may live remotely from the university campus, including overseas. Until very recently, all programs at Deakin University were required to include one unit of study available only in wholly online mode. In this format there was no face-to-face contact—all access to unit learning resources, assessment and communication between staff and students was only available via an online learning environment. The undergraduate Bachelor of Engineering program at Deakin University included the second-year engineering management

/ professional practice study unit SEB221 Managing Industrial Organisations, and it is this unit that forms the basis for the case study presented here. SEB221 consisted of four modules:

(1) Systems Concepts for Engineers and Technologists;
(2) Managing People in Organisations;
(3) Manufacturing and the Environment; and
(4) Occupational Health and Safety.

At the time of the case study presented here, SEB221 was the School's nominated wholly online unit, so regardless of the students' normal enrolment type in their other units of study, their principal form of engagement with SEB221 was via the wholly online learning environment. For the academic teaching session included in this case study, the original enrolment for SEB221 included 74 on-campus students and 58 off-campus students—132 students in total.

For students who enrol in a unit of study at Deakin University, a number of academic outcomes are possible—these are summarised in Table 1. In this case study, a student is classified as 'completing' if they did not withdraw their enrolment, as 'successful' if they achieved a grade of Pass (P) or better, and as 'unsuccessful' if they were originally enrolled but did not achieve a grade of Pass or better. For the 132 students originally enrolled in SEB221, Table 2 indicates the numbers of students achieving various unit outcomes and, where relevant, expresses these numbers as percentages of the original unit enrolment and/or the number of completing students. It can be seen that, while the percentage of completing students that were successful was very high (more than 90 percent), the percentage of the original unit enrolment that were successful was much lower (about 60 percent). It is clear that a majority of the unsuccessful students actually withdrew from the unit, rather than failed to pass per se. While some withdrawing students will be doing so for personal reasons largely beyond the influence of the university [7], an understanding of the factors that best predict failure to succeed could help in the effective targeting of scarce support resources towards those students most likely to benefit from support.

At the completion of the academic teaching period, a range of data were available for all originally enrolled students, including:

(1) final unit mark (out of 100);
(2) final unit grade (various categories as noted above);
(3) gender (female/male);
(4) normal/primary mode of enrolment (on-campus/off-campus);
(5) major course of study (SEB221 is sometimes taken by non-engineering students);
(6) age (in years);
(7) prior academic performance—measured by weighted average mark (WAM);
(8) number of LMS sessions (separate logins);
(9) total LMS session time (sum of all recorded LMS login time in decimal hours);
(10) date of first login to LMS (expressed as 'days after the commencement of the teaching period');
(11) total number of individual LMS pages viewed;
(12) total number of LMS discussion postings read; and
(13) total number of LMS discussion postings made.

Items 1–2 are outcome/output data from which 'success' status can be determined. Items 3–7 are

Table 1. Possible unit outcomes

Outcome status	Grade	Explanation
Withdrawn early	WE	Withdrawn prior to census date; no fee incurred; no mark awarded
Withdrawn late	WL	Withdrawn after census date; tuition fee payable; no mark awarded
Fail—nothing submitted	XN	Failed; no assessment submitted; mark of zero recorded
Fail	N	Failed; awarded mark less than threshold required to pass unit
Pass	P	Threshold required to pass unit $\leq$ awarded mark $\leq$ 59%
Credit	C	60% $\leq$ awarded mark $\leq$ 69%
Distinction	D	70% $\leq$ awarded mark $\leq$ 79%
High Distinction	HD	Awarded mark $\geq$ 80%

Table 2. SEB221 unit outcome statistics

Item	Calculation	Number	% of A	% of C
A. Original unit enrolment	–	132	100.0%	–
B. Number withdrawn (WE+WL)	–	46	34.9%	–
C. Number completing	A–B	86	65.1%	100.0%
D. Number failed (N+XN)	–	7	–	8.1%
E. Number successful (pass or greater)	C–D	79	59.8%	91.9%
F. Number unsuccessful (did not pass)	A–E or B+D	53	40.2%	–

time invariant data regarding students (demographic information from the university student information system). Items 8–13 are time varying data regarding student study activities (tracking data from the university LMS). This data set forms the basis of the investigation presented in this case study.

3. Methodology

All of the demographic and LMS usage data items were first screened to determine if they had any significant association with student success status. The level of statistical significance used throughout this case study is $p < 0.01$. For those data items with a significant association with student success status, a binary logistic regression was performed to determine:

- if a viable model of student success status could be developed;
- what variables, if any, were significant predictors of student success status; and
- the relative importance of any identified predictor variables.

Finally, the performance of the model was investigated to see if it could be refined to improve its predictive accuracy.

4. Results and discussion

For the categorical variables (items 3–5), a cross-tabulation with student success status was performed, and Fisher's two-sided exact test was applied. Table 3 provides a summary of the test results.

Table 3. Cross tabulations of categorical data with student success status

Data item	Fisher's two-side exact test
3. Gender	$p = 1.000$
4. Enrolment mode	$p < 0.004$
5. Course of study	$p > 0.031$

From Table 3 it can be seen that only student normal mode of enrolment has a significant association with student success status. The success rate of on-campus students was 70.5 percent, compared to 44.4 percent for off-campus students. For the continuous variables (items 6–13), an analysis of variance (ANOVA) test with student success status as the grouping variable was performed. A requirement for the ANOVA test is that the variation of the mean value of the continuous variable be similar in both grouping categories. Levene's test of homogeneity of variance can be used to test this requirement. Where Levene's test fails, a robust ANOVA test using the Welch test statistic can be performed instead. Table 4 provides a summary of the test results.

From Table 4 it can be seen that all variables except age were significantly associated with student success status. However, a number of the variables related to LMS usage are essentially cumulative tallies (items 8, 9, 11, 12 and 13), only becoming validly available at the end of the teaching period, and hence, in this case, offer limited predictive ability for timely interventions that might be made to improve student academic performance. So, the data items that have a significant association with student success status and which may offer practical predictive utility are: (4) enrolment mode, (7) prior academic performance (WAM) and (10) date of first login to LMS.

These three variables, their three two-way cross products and their single three-way cross product were initially used as the predictor variables in a binary logistic regression with student success status as the dependent variable. The resultant logistic regression model included all three variables plus the cross product of mode-by-WAM. However, enrolment mode was not a significant variable ($p > 0.056$). So, the logistic regression was repeated using only WAM, date of first login to LMS and mode-by-WAM as predictors. Table 5 shows a summary of the resultant binary logistic regression model.

All of the predictor variables were significant. A constant term (β_0 in Eq. 2) was required for the

Table 4. ANOVA tests of continuous data against student outcome status

Data item	Successful mean	Unsuccessful mean	Levene's test	ANOVA test	Robust ANOVA test
6. Age	25.9 years	28.4 years	$p > 0.256$	$p > 0.119$	–
7. WAM	66.07%	54.40%	$p < 0.003$	–	$p < 0.002$
8. No. LMS sessions	49.29	5.45	$p < 3.5 \times 10^{-6}$	–	$p < 2.2 \times 10^{-16}$
9. Total LMS time	18.53 hrs	1.87 hrs	$p < 4.4 \times 10^{-10}$	–	$p < 2.9 \times 10^{-15}$
10. Date of 1st login to LMS*	–4.39 days	36.51 days	$p < 2.7 \times 10^{-4}$	–	$p < 2.3 \times 10^{-7}$
11. LMS pages viewed	119.2 pgs	14.6 pgs	$p < 3.6 \times 10^{-9}$	–	$p < 8.4 \times 10^{-17}$
12. LMS posts read	565.4 msg	45.0 msg	$p < 0.004$	–	$p < 0.008$
13. LMS posts made	8.77 msg	0.64 msg	$p < 7.8 \times 10^{-6}$	–	$p < 5.8 \times 10^{-13}$

* Expressed as days after the commencement of the teaching period.

Table 5. Summary of first binary logistic regression model

Predictor variable	β	Significance	e^{β}
Date of 1st login to LMS	0.039	$p < 0.0010$	1.040
WAM	–0.086	$p < 0.0003$	0.918
Mode-by-WAM	0.029	$p < 0.0004$	1.029
Constant	1.926	$p > 0.076$	6.862

model to give predicted values between 0 and 1—in this case, a 1 represents a prediction of a student not succeeding with a final unit grade of pass or better. As with other forms of regression modelling, there are some statistics that can be calculated to test the performance of the model. The Hosmer and Lemeshow Test provides a measure of the 'goodness of fit' of the model; p values of > 0.05 are generally considered to indicate satisfactory goodness of fit. For the model in Table 5 the Hosmer and Lemeshow test statistic was $p > 0.50$, so the model has good fit properties. Another commonly used statistic is the coefficient of determination, which indicates the proportion of the variance in the data accounted for in the model. In the case of linear regression, this is the R^2 statistic. An analogous statistic for logistic regression is the Nagelkerke R^2 statistic; larger values, up to the maximum value of 1, indicate increasing goodness of fit. For the model in Table 5 the Nagelkerke R^2 statistic was 0.509, indicating again that the model has good fit properties.

For the development of possible actions to improve student success, it is important to understand the regression model obtained. The variables x_1, x_2 and x_3 in Eq. 2 are the predictor variables listed in Table 5. The coefficients β_1, β_2 and β_3 in Eq. 2 are the β values given in Table 5 associated with each predictor variable—as noted above, the constant term is β_0 in Eq. 2. Due to the nature of Eq. 1, the larger the positive value of each term in Eq. 2, that is, the larger the positive value for each $\beta_n x_n$ product, the larger the result predictor value for $f(z)$, and the more likely the regression model is to predict a particular student as unsuccessful. For the significant model predictor variables of date of 1st login to LMS and WAM, the associated β coefficient values make intuitive sense. Date of 1st login to LMS is expressed as days after the commencement of the teaching period, so the longer a student delays accessing the LMS, coupled with a positive β coefficient, the less likely they are to succeed. The larger the value of WAM, the stronger the student's prior academic performance, and coupled with a negative β coefficient, the higher the WAM, the more likely they are to succeed. The direct impact of these predictor variables can be seen by the differences in their mean values for successful and unsuccessful students given in Table 4. Successful students have, on average, a significantly higher WAM score, and, on average, their date of first access of the LMS is much earlier—the negative mean value in Table 4 indicating an access date prior to the commencement of the formal teaching period. The Mode-by-WAM cross product as a predictor in the model is less obvious, but it suggests that mode of study is important. Mode here refers to the student's normal mode of enrolment, as for this unit, all students were studying in 'wholly online' mode. In the original regression data set, on-campus mode was given a data value of 1, and off-campus mode was given a data value of 2. So, there is evidence that off-campus students are less likely to succeed. As noted in Table 3, mode of study did have a significant association with student success status, and off-campus students were less likely to succeed.

In the regression model presented in Table 5, a measure of the effect size of each predictor variable is given by e^{β}—the exponential of the β coefficient is required because of the form of Eq. 1. The range of e^{β} is small; from 0.918 to 1.04—indicating that all predictor variables are approximately equally important. Table 6 summarises the predictive performance of the regression model given in Table 5, based on a binary cut value of 0.5.

While the overall accuracy of prediction is 78.6 percent, the performance of predicting unsuccessful students (55.3 percent) is significantly lower than the very good performance of predicting successful students (92.4 percent). The original data set used in this case study contains a small number of students without a valid WAM recorded. This can occur when a student transfers into the engineering program at Deakin University with advanced standing that sees them enrolled in SEB221 in their first semester of study at Deakin University. Given that WAM turned out to be an important predictor variable in the model above, the absence of a WAM value means that the first model is unable to predict a success status for these students. To consider this situation, the other two of the three originally identified useful data items, (4) enrolment mode and (10) date of first login to LMS, and their two-way cross product were used as the predictor variables in a binary logistic regression with student success status as the dependent variable. The resultant logistic regression model included both vari-

Table 6. Predictive performance of first binary logistic regression model

	Predicted		
Observed	**Successful**	**Unsuccessful**	**% Correct**
Successful	73	6	92.4%
Unsuccessful	21	26	55.3%
Overall%			78.6%

Table 7. Summary of second binary logistic regression model

Predictor variable	β	Significance	e^{β}
Date of 1st login to LMS	0.046	$p < 0.0002$	1.047
Mode	1.433	$p < 0.0015$	4.193
Constant	–2.783	$p < 0.0001$	0.062

Table 8. Predictive performance of second binary logistic regression model

	Predicted		
Observed	Successful	Unsuccessful	% Correct
Successful	74	5	93.7%
Unsuccessful	25	28	52.8%
Overall%			77.3%

ables, but not their cross product. Table 7 shows a summary of the binary logistic regression model.

Both of the predictor variables were significant. For the model in Table 7 the Hosmer and Lemeshow test statistic was $p > 0.14$, and the Nagelkerke R^2 statistic was 0.437, indicating that the model has good fit properties. Table 8 summarises the predictive performance of the regression model given in Table 7, based on a binary cut value of 0.5.

The overall predictive accuracy of both models is similar, and similar to other model predictive accuracies reported in the literature [1–2, 13–14]. As for the first model in Table 5, the β coefficients of the predictor variables in the second model are positive, indicating that the later a student first accesses the LMS and whether they are normally enrolled in off-campus mode are predictors of not attaining a final unit grade of pass or better.

While the overall accuracy of prediction of the second model is similar to the first model, it has the advantage of being able to provide a prediction for students who do not have a WAM score. In both Table 5 and Table 7 it can be seen that the proportion of successful students incorrectly predicted as unsuccessful is low, while the proportion of unsuccessful students incorrectly predicted as successful is significant. Given that the focus of this investigation is the identification of students 'at risk' academically, there is scope for trading off some accuracy in the successful prediction rate if it were possible to improve the unsuccessful prediction rate. The literature suggests that the prediction accuracy of student success modelling can be improved by combining (fusing) the prediction results from more than one model (model ensemble) [1]. In particular, the decision scheme that was found to be most accurate in one investigation was one based on identifying a student as at risk if at least one of the ensemble models identified that student as being at risk [14]. This 'logical OR' decision scheme was applied to the prediction outputs from both previous regression models, and the resulting performance of the combined model is given in Table 9.

Table 9. Predictive performance of combined binary logistic regression model

	Predicted		
Observed	Successful	Unsuccessful	% Correct
Successful	72	7	91.1%
Unsuccessful	21	32	60.4%
Overall%			78.8%

It can be seen that the overall prediction performance of the combined model is only marginally better than the individual models, and while the combined correct prediction rate for successful students has declined slightly, the combined correction prediction rate for unsuccessful students has increased by 9.2 percent over the first model and by 14.4 percent over the second model. The combined model is significantly better at predicting students not successfully completing the unit.

The case study described here doesn't offer a precise model that can be literally interpreted as the formula for predicting student academic performance, but it does suggest those factors from the available data set that provide an advance indicator of whether a particular student might be at risk of not succeeding. Both constituents of the combined model included time-invariant (WAM and/or mode of study) and time-variant data (date of 1st login to LMS); reinforcing this as a desirable feature of such models. It is unlikely that institutions can ever stop all students from withdrawing or failing—some factors contributing to these outcomes (in this case, WAM and mode of study) are essentially beyond the control of the universities. However, it was found here that date of 1st login to LMS was a significant predictor of success—student access to the LMS is easily monitored, and any detected delays in initial access to the system could be followed up with a contact to see how such students are going. The factors contributing to student academic performance are likely to be at least partially context-dependent [2, 7–8, 10, 13, 15–16], and the specific characteristics of the student group investigated here (second-year engineering students studying in wholly online mode) may limit the generalizability of the findings. In this case study, the proportion of completing students attaining a 'fail' grade was quite small, making it impossible to model the factors specifically contributing to that outcome. Given that failing students are of great interest to educators, a larger student cohort/data set would be required for analysis, perhaps the enrolment for an entire year level, or an entire program. The performance of data-driven models for prediction depend largely on the size and quality of the data sets representing the phenomenon under investigation [1, 14], so the relatively small data set (in terms of both number of variables and number of student cases) available for analysis in the case study presented here is a significant limitation. The analysis presented is likely to be enhanced by the inclusion of more candidate predictor variables from institutional databases containing student data, and from the inclusion of data from more students.

5. Conclusion

This paper presents an academic analytics investigation into the modelling of academic performance of engineering students enrolled in a second-year class. The modelling method used was binary logistic regression, and the target predicted variable was 'success status'—defined as those students from the total originally enrolled group that achieved a final unit grade of pass or better. From the data available for modelling, the significant predictor variables included mode of study and date of 1st login to LMS, and for students that had a measure of prior academic performance recorded, weighted average mark (WAM) was also a significant predictor. These results confirm findings by others that prior academic performance is an important predictor of current performance, and that time variant data, in addition to time-invariant data, can improve the performance of student predictive models. Two models were developed, one for use when WAM was available, and one for use when WAM was not available. A combined model using results from both individual models was found to have superior predictive performance, again confirming findings by others. This paper show that student data stored in institutional systems can be used to predict student academic performance with reasonable accuracy, and it provides one relatively simple but effective methodology for achieving this. Importantly, significant predictor variables are identified that offer the ability to develop targeted and timely interventions to proactively improve student success and retention outcomes. While much of the literature on university student retention and progression focuses on the first year of university study, the case study presented here addresses the lesser explored, yet still fundamentally important, issue of student academic performance in the second year of study. The case study presented here is modest in scope, and could be enhanced by the inclusion of more candidate predictor variables from institutional databases containing student data, and from the inclusion of data from more students.

References

1. D. Delen, A comparative analysis of machine learning techniques for student retention management, *Decision Support Systems*, **49**(4), 2010, pp. 498–506.
2. C. H. Yu, S. DiGangi, A. Jannasch-Pennell and C. Kaprolet, A Data Mining Approach for Identifying Predictors of Student Retention from Sophomore to Junior Year, *Journal of Data Science*, **8**(2), 2010, pp. 307–325.
3. D. W. Knight, L. E. Carlson and J. F. Sullivan, Improving Engineering Student Retention through Hands-On, Team Based, First-Year Design Projects, in *31st International Conference on Research in Engineering Education*, American Society for Engineering Education, Honolulu, 2007, pp. 11–13.
4. G. Ragusa and C. T. Lee, The impact of focused degree projects in chemical engineering education on students' research performance, retention, and efficacy, *Education for Chemical Engineers*, **7**(3), 2012, pp. e69–e77.
5. R. Jain, B. Shanahan and C. Roe, Broadening the Appeal of Engineering—Addressing Factors Contributing to Low Appeal and High Attrition, *International Journal of Engineering Education*, **25**(3), 2009, pp. 405–418.
6. V. Tinto, From Theory to Action: Exploring the Institutional Conditions for Student Retention, in *Higher Education: Handbook of Theory and Research,* Vol. 25, Springer, New York, 2010, pp. 51–89.
7. M. W. Ohland, S. D. Sheppard, G. Lichtenstein, O. Eris, D. Chachra and R. A. Layton, Persistence, Engagement, and Migration in Engineering Programs, *Journal of Engineering Education*, **97**(3), 2004, pp. 259–278.
8. G. Zhang, T. J. Anderson, M. W. Ohland and B. J. Thorndyke, Identifying Factors Influencing Engineering Student Graduation: A Longitudinal and Cross-Institutional Study, *Journal of Engineering Education*, **93**(4), 2004, pp. 313–320.
9. R. Reason, D., Student Variables that Predict Retention: Recent Research and New Developments, *Journal of Student Affairs Research and Practice*, **46**(3), 2009, pp. 482–501.
10. J. P. Campbell and D. G. Oblinger, *Academic Analytics*, EDUCAUSE, Boulder, 2007.
11. A. Caison, Analysis of institutionally specific retention research: A comparison between survey and institutional database methods, *Research in Higher Education*, **48**(4), 2007, pp. 435–451.
12. S. Lonn and S. D. Teasley, Saving time or innovating practice: Investigating perceptions and uses of Learning Management Systems, *Computers & Education*, **53**(3), 2009, pp. 686–694.
13. L. P. Macfadyen and S. Dawson, Mining LMS data to develop an 'early warning system' for educators: A proof of concept, *Computers & Education*, **54**(2), 2010, pp. 588–599.
14. I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis and V. Loumos, Dropout prediction in e-learning courses through the combination of machine learning techniques, *Computers & Education*, **53**(3), 2009, pp. 950–965.
15. A. F. Cabrera, A. Nora and M. B. Castaneda, College Persistence: Structural Equations Modeling Test of an Integrated Model of Student Retention, *The Journal of Higher Education*, **64**(2), 1993, pp. 123–139.
16. C. M. Vogt, Faculty as a Critical Juncture in Student Retention and Performance in Engineering Programs, *Journal of Engineering Education*, **97**(1), 2008, pp. 27–36.
17. C. Hart, Factors Associated With Student Persistence in an Online Program of Study: A Review of the Literature, *Journal of Interactive Online Learning*, **11**(1), 2012, pp. 19–42.

Stuart Palmer is an Associate Professor in the Deakin University Faculty of Science and Technology. Prior to that, he was an academic staff member of the School of Engineering at Deakin University for 12 years. He is an active researcher in the area of frequency domain analysis and engineering education.