Predicting success: How learners' prior knowledge, skills and activities predict MOOC performance

Gregor Kennedy
Centre for the Study of Higher Education
The University of Melbourne
gek@unimelb.edu.au

Carleton Coffrin
National ICT Australia
Victoria Research Laboratory Melbourne, Victoria, Australia
carleton.coffrin@nicta.com.au

Paula de Barba,
Centre for the Study of Higher Education
The University of Melbourne
paula.de@unimelb.edu.au

Linda Corrin
Centre for the Study of Higher Education
The University of Melbourne
l.corrin@unimelb.edu.au

ABSTRACT
While MOOCs have taken the world by storm, questions remain about their pedagogical value and high rates of attrition. In this paper we argue that MOOCs which have open entry and open curriculum structures, place pressure on learners to not only have the requisite knowledge and skills to complete the course, but also the skills to traverse the course in adaptive ways that lead to success. The empirical study presented in the paper investigated the degree to which students' prior knowledge and skills, and their engagement with the MOOC as measured through learning analytics, predict end-of-MOOC performance. The findings indicate that prior knowledge is the most significant predictor of MOOC success followed by students' ability to revise and revisit their previous work.

Categories and Subject Descriptors
K.3.1 [Computer Uses in Education]: Distance learning—MOOC; J.1 [Administrative Data Processing]: Education

General Terms
Measurement, Performance, Experimentation, Human Factors.

Keywords
Prior knowledge, learning analytics, engagement, MOOCs

1. INTRODUCTION
The last five years have seen a rapid rise in the popularity of Massive Open Online Courses, or MOOCs. This rise in popularity has been reflected both in the number of learners enrolling in these courses and the number of universities now offering courses in this format. While there are many challenges in providing learners with high quality educational experiences at such a large scale, MOOCs have created significant opportunities for educational researchers to better understand how learners develop their knowledge and understanding through online learning. The sheer numbers of learners who participate in MOOCs – often in the thousands – means that researchers have access to large datasets of each learners' online interactions which, through the use of learning analytics, can be used to develop a greater understanding of learners' online experiences, processes, and outcomes [1].

While retention and pass rates have often been used as markers of success in traditional courses, these are more problematic metrics in MOOCs (see [2][3]). As MOOCs are free, students have less financial incentive to persist with the course and this may be a reason for the high attrition rates. Moreover, is it possible given MOOCs have open enrollments and do not require any demonstration of pre-existing experience, qualifications or credentials, that enrolling students who have poor prior knowledge and skills may find it difficult to successfully engage with and complete the course.

1.1 MOOC Preparedness: Students' Prior Knowledge and Skills
For many years, educational theories and frameworks have implicated students' prior knowledge as a key ingredient in an individuals learning success. Piaget's [4] contention that the development of understanding is through a process of assimilation and accommodation is underpinned by the idea of pre-existing or prior knowledge. These foundational concepts in educational psychology suggest that a student's understanding is developed by building on and modifying his or her existing knowledge structures – or schema. As [5] suggest, humans "come to formal education with a range of prior knowledge, skills, beliefs, and concepts that significantly influence what they notice about the environments and how they organize and interpret it" (p. 10).

In many ways the emphasis of developmental and cognitive psychology on the importance of prior knowledge underpins the constructivist approaches to teaching and learning which currently dominate the learning technology landscape [6]. These constructivist approaches emphasise the need to understand what individual students bring to each learning situation in terms of their background knowledge and understanding. The most beneficial teaching and learning environments, rather than having teachers simply broadcast information for students to learn, take into account the potentially very different starting points of individual students. The effectiveness of any particular instructional situation or environment is dependent on accounting for different student perspectives, backgrounds and prior knowledge.

As noted by Bransford, Brown and Cocking [5] students not only possess prior knowledge in specific discipline-based content areas, they

LAK '15, March 16 - 20, 2015, Poughkeepsie, NY, USA Copyright is held by the owner/author(s). Publication rights licensed to ACM. ACM 978-1-4503-3417-4/15/03…$15.00
http://dx.doi.org/10.1145/2723576.2723593

also have pre-existing generic learning skills, which they bring to learning situations. Education research is replete with taxonomies of the types of cognitive or learning skills and strategies that students draw upon in teaching and learning environments, including problem solving, critical thinking, self-regulation and metacognition [7][8]. Students' ability to further develop and employ these skills effectively in the course of their learning has received a great deal of attention by education researchers [9].

In short, previous educational research has established that prior knowledge and skills – both in terms of content knowledge and generic learning skills, such as problem solving – can greatly influence students' learning success.

1.2 MOOC Preparedness: Navigating Open Curricula

A feature of the emerging MOOC landscape has been the opportunity for educators to experiment with curriculum structures. While there is some debate about the origins of MOOCs, many point to George Siemens and Stephen Downes' course on Connectivism and Connective Knowledge as one of the first [3]. This course was founded on a connectivist pedagogical framework [10], which at its heart advocates a peer-based, networked approach to learning. While the "open" in MOOCS is often taken to refer both to "open access" (anyone with access to the internet can enroll) and "open educational resources" (depending on the course, students make use of freely available resources on the Internet), it can also refer to "open" curriculum structures that are consistent with a connectivist philosophy of learning.

What this means in practice is that, unlike more traditional courses in which students' engagement with content is structured around weekly topics and assignments which are followed in a progressive, linear way, connectivist or "cMOOCs" with a more open curriculum structure are less proscriptive and give students greater flexibility in the ways in which they can engage with the course, other learners and the curriculum material.

A clear implication of this is that students who choose to participate in a MOOC, particularly those with an open curriculum structure, need to be prepared for a certain amount of self-direction. With cMOOCs there is arguably a greater onus on students to manage and plan their own learning as they choose what aspects of the course they want and need to engage with, based on their own interest and prior knowledge and skills. Moreover, with open curriculum structures in which students can complete tasks and assignments in any order, there is an opportunity for learners to revisit and revise in ways that are often not advocated or even available in more structured linear courses.

Set against this background, this paper presents an investigation of how students' preparedness for MOOCs impacts on their participation and success in the course. More specifically, this paper considers how students' prior content knowledge in an advanced area of computer science (Discrete Optimization) and their problem solving skills related to this area of computer science impact on their patterns of engagement as measured through learning analytics and their learning outcomes or performance.

2. METHOD

2.1 Course Structure and Participants

Data were collected from the first session of Discrete Optimization, a MOOC provided by the University of Melbourne on the Coursera platform. Discrete Optimization was first offered in June 2013. It was a graduate level course, which assumed incoming learners had a background in computer science and strong computer programming skills. It consisted of nine weeks of material presented in an open curriculum structure. That is, all of the assignments and lectures were made available in the first week, and while there was a logical sequence and order in which material and assignments could be covered broadly based on difficulty, learners designed their own study plan which they completed at their own pace (see [2] [11] for further details).

The inaugural session of this course attracted the interest of 37,777 learners, with 22,731 starting the course, 6,635 active in the assignments, and 774 receiving a certificate of completion for the course. The sample for this investigation was drawn from a subset of students who started the course: those who were active in the assignments (n=6,635), and those who received a certificate of completion (n=774). These groups were not mutually exclusive.

The assessments in Discrete Optimization consisted of seven programming assignments designed around problem solving tasks. Each of the core assignments presented the learner with an emulation of what could be regarded as a real world Discrete Optimization experience; that is, an employer telling them: "solve this problem, I don't care how". The lecture materials contained the necessary concepts and ideas to solve the assignments, but the most appropriate technique to apply was left for the learners to work out. This assignment design helps to prepare learners for how optimization is conducted in the real world and has been used effectively over a number of years in the classroom version of Discrete Optimization [11].

The programming assignments increased in difficulty, with each assignment requiring students to show a deeper understanding of the course material. For example, the "first" (i.e. least difficult) assignment, Screen Name, was very simple and only required a basic understanding of computer programming to complete. The second assignment, Knapsack, was more challenging and required significant prior content knowledge in computer science. The next assignment, Graph Coloring, required all of the skills of the previous assignments – basic computer programming and requisite computer science content knowledge – but in order to complete this assignment students needed to have more sophisticated problem-solving skills. The remaining assignments increased in difficulty and required students to show progressively advanced knowledge and skills in computer programming, computer science, and problem solving. Furthermore, these remaining assignments were designed to be very challenging. A student striving to achieve full marks needs to master nearly all of the course material. By this design, it is expected that high achieving students will revisit past assignments and revise their work as they develop their skills in the subject area.

2.2 Learning Analytic Data Acquisition

All data analysed in this study were collected automatically by the Coursera platform. The Coursera platform records a vast amount of information on learners' activities, and provides end users (instructors, administrators, researchers) with three general views: course-wide statistics which provide an aggregated overview of activity for the entire class; a grade book which provides a summary information about each learner's performance; and an event log, which tracks every interaction the learners have with the platform. The grade book and the event log data formed the basis for the learning analytics used in this investigation. The grade book data were exported as an excel spreadsheet, and the event log data were provided as an SQL database for which custom SQL queries and python scripts were used to extract, compute, and aggregate the required student metrics. These two data sets were merged into a single data file before being imported into SPSS version 21 for analysis.

2.3 Measures

The measures used in this investigation are defined below.

2.3.1 Knapsack Points

Knapsack Points was derived from the knapsack programming assignment, which assessed students' prior content knowledge in computer science. Students who have a strong background in computer science would typically have been exposed to a method of algorithm design called "dynamic programming". The knapsack problem is routinely used in computer science teaching to assess dynamic programming, and completing this problem well is indicative of a strong background in computer science. Students could score between 0 to 60 points on the knapsack problem, and *Knapsack points* was a measure of the total points earned on the knapsack assignment on the last day of class.

2.3.2 Graph Coloring Points

Graph Coloring Points was derived from the graph coloring programming assignment, which assessed students' skills in computer science problem solving. Students could score between 0 to 60 points on this assignments and *Graph Coloring Points* was a measure of the total points earned on the graph coloring assignment on the last day of class.

2.3.3 Assignment Submissions

Assignment Submissions was a measure of the total number of times a student submitted any assignment during the course. This value can range from 0 to infinity, however if a student submitted each assignment just once he or she would have an *Assignment Submissions* value of 37 (reflecting independent submissions of sub-components of each of the seven assignments). The assignment submissions value was calculated using a simple frequency count from each students' event log data.

2.3.4 Active Days

Active Days was a measure of the total number of days a student was actively submitting assignments in the course. As the course was nine weeks long, this value could range from 0 to 62. Active days was calculated using the event log data by taking the timestamp of each learner's first assignment submission and subtracting it from the timestamp of the last assignment submission, and rounding down to a whole day.

2.3.5 Assignment Switches

Assignment switches was a measure of the number of times a learner switched from submitting one assignment to a different assignment. This measure reflects the degree to which students were following a more traditional linear progression in the course. Students with a higher score on *Assignment Switches* were more inclined to move between assignments and revisit assignments that they had previously completed. The value for *Assignment Switches* could range from 0 to infinity, however, a learner who worked on the assignments in a linear order and did not revisit any previous assignments (i.e. moved from one assignment to another in a linear sequence) would have a value of 6. The assignment switches value was calculated by parsing the learner's event log data in chronological order. Each time the assignment submission type changed the switching value is increased by one.

2.3.6 Total Points

Total Points was a measure of the students' overall performance in the course and was calculated as the cumulative points earned by the learner across all assignments on the final day of the course. The value range for this measure was 0 to 396.

3. Results

The analyses conducted as part of this investigation employed two samples. The first sample included all participants (n=6,635) who were active in the MOOC as determined by submitting at least one assignment. The second sample – a subset of the first – only included participants who passed the course (n=774). These two samples were investigated as it was expected that, given the focus on prior knowledge and skills, distinct patterns may emerge for those who passed the course and those who did not.

3.1 All Participants

Descriptive statistics for all variables and the correlations between them for all 6,635 participants are presented on Table 1. Correlations between all variables were significant. Strong positive correlations were seen among knapsack points, graph coloring, active days, and assignment switching, particularly between graph coloring and active days and assignment switching, and between active days and assignment switching. While positive, assignment submission was weakly correlated with all other variables.

A stepwise multiple regression was conducted to determine the degree to which these variables were able to predict total points. Multicollinearity was checked and was well within accepted parameters. At step 1 of the analysis knapsack points was entered into the regression model and was significantly related to total points, $F(1,6633) = 7401.77$, $p<.001$. This model accounted for approximately 53% of the variance of total points (Adj. $R2 = .527$). At step 2 of the analysis graph coloring was entered into the model and was also significantly related to total points, $F(2,6632) = 15982.75$, $p<.001$. This model accounted for approximately 83% of the variance of total points (Adj. $R2 = .828$). Total points was primarily predicted by graph coloring, and to a lesser extent by knapsack points. Finally, at step 3 of the analysis the three remaining variables were entered into the regression model and active days and assignment switching were statistically significant, $F(5,6629) = 13044.39$, $p < .001$. This model accounted for approximately 91% of the variance of total points (Adj. $R2 = .908$). Total points were primarily predicted by graph coloring, assignment switches, and active days, and to a lesser extent by knapsack points. Assignment submission's contribution to the model was not significant. Regression coefficients and other relevant statistics for each model are presented in Table 2.

3.2 "Passing" Participants

From the 6,635 learners who participated in the previous analysis, 774 passed the course. Descriptive statistics and correlations between variables for these participants are presented in Table 3. There was a moderate positive correlation between graph coloring and knapsack points, and assignment submission and assignment switching. There was a weak positive correlation between assignment submissions and both graph coloring and active days; and assignment switching with graph coloring and active days.

Again, a stepwise multiple regression was used to determine the degree to which variables predicted total points (tests of multicollinearity were again acceptable). At step 1 of the analysis knapsack points was entered into the regression equation and was significantly related to total points, $F(1,772) = 102.87$, $p<.001$; and accounted for only about 12% of the variance in total points (Adj. $R2 = .116$). The second step of the model, in which graph coloring was entered, was significant ($F(2,771) = 262.56$, $p<.001$) and accounted for approximately 40% of the variance of total points (Adj. R2 =

Table 1. Descriptive statistics and correlation matrix for the primary study variables – All Cases (n = 6,635)

	M	SD	1	2	3	4	5	6
1. Knapsack points	27.94	25.74		.658**	.033**	.644**	.527**	.726**
2. Graph coloring	11.94	20.02			.054**	.809**	.718**	.891**
3. Assignment submissions	69.94	2340.79				.051**	.030**	.042**
4. Active days	11.51	17.74					.738**	.860**
5. Assignment switching	2.31	4.43						.833**
6. Total points	66.41	94.79						

Notes. ** $p < .001$.

Table 2. Stepwise Regression Results – All Cases (n = 6,635)

	Model	b	SE-b	Beta	Pearson r	sr^2	Struct. Coef.
1	Constant	-8.328	1.181				
	Knapsack points**	2.675	.031	.726	.726	.527	1.000
2	Constant	-.193	.716				
	Knapsack points**	.910	.025	.247	.726	.035	.038
	Graph coloring**	3.448	.032	.728	.891	.300	.330
3	Constant	-1.564	.525				
	Knapsack points**	.651	.019	.177	.726	.017	.762
	Graph coloring**	1.819	.033	.384	.891	.042	.935
	Assign. sub.	.000	.000	-.005	.042	.002	.044
	Active days**	1.090	.038	.204	.860	.012	.902
	Assing. switch.**	6.728	.124	.314	.833	.041	.874

Notes. The dependent variable was total points. sr^2 is the squared semi-partial correlation. Assign. subm. = Assignment submissions. Assign. switch. = Assignment switching. Struct. Coef. = Structure Coefficient. ** $p < .001$.

.404). Total points were primarily predicted by graph coloring, and to a lesser extent by knapsack points. Finally, at step 3 of the analysis assignment switching, assignment submission and active days were entered into the regression model which was statistically significant, $F(5,768) = 121.67$, $p < .001$. This model accounted for approximately 44% of the variance of total points (Adj. $R2 = .438$). Total points were primarily predicted by graph coloring, and to a lesser extent by assignment switching and knapsack points. The contributions of assignment submissions and active days to this model were not significant. Regression coefficients and other relevant statistics for each model are presented in Table 4.

4. DISCUSSION

This investigation considered how students' prior knowledge in computer science and problem solving, and their engagement within an open MOOC curriculum, impacted on their MOOC performance.

The stepwise regression analyses conducted with all participants clearly showed that prior knowledge in computer science was a key indicator of success, but that the impact of prior knowledge of the content area was suppressed somewhat by prior problem solving skills when it was introduced into the regression model. These findings suggest that more generic prior knowledge in problem solving skills is more important to students' success than prior knowledge in the content area.

The stepwise regression model indicated that while both these forms of prior knowledge maintained their importance when participation in the MOOC was considered, students' propensity to exploit the open curriculum structure by switching between assignment tasks in a non-linear fashion was also important. The full model also indicated that regular or persistent activity in the class was significantly related to successful performance.

A second stepwise regression analysis was conducted to determine whether the pattern of associations found in the first regression model with all active students could be replicated with just those students who passed the course. The results from this second analysis and sample showed a pattern of results that was similar but different in two primary respects from the first regression analysis. The analyses undertaken with the second sample was similar in that prior content knowledge and problem solving skills still significantly predicted students' performance, with problem solving skills still showing a stronger effect than prior content knowledge. In addition, assignment switching remained a significant predictor of success. However, in the second set of regression analyses, the number of active days students spent on the course was not a significant predictor, and the overall amount of variance in the outcome explained was markedly lower than that seen in the first model.

Prior knowledge – both content and problem solving skills – had a very strong association with students' performance, particularly with the sample of all active students; prior knowledge variables alone accounted for 83% of the variance in students' performance. When compared to the sample of "passing" students this was approximately double the amount of variance explained in performance. This is reflected in the full regression models for both samples: the proportion of variance explained in students' performance in "passing" students sample was markedly lower than that of the sample of all active students (44% and 91% respectively).

A clear possible explanation for this finding is that measures of prior knowledge are strong predictors of success in the cohort of "active" students because this sample contains a larger number of learners who attempt an assignment or two, but recognize they do not have the prior knowledge and skills to complete the course, subsequently disengage, and ultimately drop-out. This would result in a large number of learners being included in the sample with a low total points score (thereby reducing the overall variance of this measure). If a large cohort of students among the 6,635 are in this category it would explain why prior knowledge measures account for such a high proportion of variance in the outcome.

However, it is important to note that this explanation does not diminish the influential role of prior knowledge in predicting students' performance in the MOOC. The regression analysis undertaken with the second sample, in which the variance in total points is higher, shows that prior knowledge and skills are still very strong predictors of variance in students' learning outcomes.

While prior knowledge is an important variable in predicting success in the "passing students" sample, there is a significant proportion of variance left unexplained in this model. That is, additional factors must be contributing to students' success in the MOOC. It is reasonable to conclude that prior knowledge and skills are a necessary but not sufficient condition in predicting students' MOOC completion and success.

The results presented indicate that after measures of prior knowledge, students' ability to exploit the open curriculum structure was a significant factor in their ultimate success. Students who were inclined to return to, revisit and revise their assignments – encouraged by the design of the learning tasks and the open curriculum structure – were more likely to perform well.

Interestingly, while *active days* had a significant association with performance for the sample of all participants, it was not predictive of success for those students who passed the course. While the argument based on the difference between the two samples may go some way in explaining this, it is interesting to reflect on why *active days* was not a

Table 3. Descriptive statistics and correlation matrix for the primary study variables – Only Passed Cases (n = 774)

	M	SD	1	2	3	4	5	6
1. Knapsack points	58.93	2.98		.373**	-.054	.014	.005	.343**
2. Graph coloring	50.57	7.78			.078*	.033	.154**	.625**
3. Assignment submissions	206.45	154.40				.085*	.373**	.108**
4. Active days	47.23	11.64					.120**	.062
5. Assignment switching	11.76	6.87						.278**
6. Total points	295.50	47.67						

Notes. ** $p < .001$, * $p < .05$.

Table 4. *Stepwise Regression Results – Only Passed Cases (n = 774)*

	Model	b	SE-b	Beta	Pearson r	sr^2	Struct. Coef.
1	Constant	-27.917	31.928				
	Knapsack points**	5.488	.541	.343	.343	.118	1.000
2	Constant	-3.690	26.261				
	Knapsack points**	2.039	.479	127	.343	.014	.538
	Graph coloring**	3.540	.183	.578	.625	.287	.981
3	Constant	-23.409	26.118				
	Knapsack points**	2.226	.467	.139	.343	.016	.192
	Graph coloring**	3.329	.181	.543	.625	.247	.747
	Assign. sub.	.000	.009	.000	.108	.000	.000
	Active days	.079	.111	.019	.062	.000	.029
	Assing. switch.**	1.327	.204	.191	.278	.031	.263

Notes. The dependent variable was total points. sr^2 is the squared semi-partial correlation. Assign. subm. = Assignment submissions. Assing. switch. = Assignment switching. Struct. Coef. = Structure Coefficient. ** $p < .001$.

significant predictor for passing students. It seems likely that students who passed the course were completing it at different rates – some students finished the course quickly, in as few as two weeks, while others used the full nine weeks. This seems to be a possible explanation for why "degree of activity" did not predict passing students' end-of-course performance.

The descriptive statistics indicate learners are submitting each assignment many times more than would have been expected in a linear course design (4.7 submissions of each assignment, on average). However, the findings showed that this high rate of assignment submission did not impact on students' performance in the course, for either sample. It seems that it is not the number of submissions, but general activity or engagement in the course, and more importantly, the degree to which the learner is prepared to move between assignments within the course that is associated with course success.

A clear implication of the findings from this paper is that it may be useful early on in a MOOC to provide students with diagnostic measures of prior content knowledge and learning skills such as problem solving, as this would provide learners with an indication of their pre-existing competency to complete and succeed in the course. Such measures would be not only useful for the student, they would also be useful for staff who are teaching the course, both in terms of setting expectations, and determining which students are encountering difficulty early on. As we have suggested in our earlier work [2], such measures may provide scope for teaching staff and MOOC developers to intervene early in a students' engagement with a MOOC and direct them to alternative and/or supplementary learning resources.

5. ACKNOWLEDGMENTS

The authors acknowledge the support of the Learning Analytics Research Group at the University of Melbourne, and NICTA which is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program.

6. REFERENCES

[1] Siemens, G., and Long, P. 2011. Penetrating the fog: Analytics in learning and education. *Educause Review*, 46(5), 30-32.

[2] Coffrin, C., Corrin, L., de Barba, P., and Kennedy, G. 2014. Visualizing patterns of student engagement and performance in MOOCs. In *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge*. ACM, New York, NY, 83-92.

[3] Daniel, J. 2012. Making sense of MOOCs: Musings in a maze of myth, paradox and possibility. *Journal of Int. Media in Education*, 3.

[4] Piaget, J. 1973. *To understand is to invent: The future of education.* Grossman, New York, NY.

[5] Bransford, J. D., Brown, A. L., Cocking, R. R. 1999. *How people learn: Brain, mind, experience, and school*. Nat. Academy Press.

[6] Hawkins, D. (1994). Constructivism: Some history. In P.J. Fensham, R.F. Gunstone & R.T. White (Eds), *The content of science: A constructivist approach to its teaching and learning* (pp. 9-13). Falmer, London, UK.

[7] Pintrich, P. R., Smith, D., García, T., and McKeachie, W. 1991. *A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ).* Ann Arbor, Michigan, USA.

[8] Weinstein, C. E. and Mayer, R. E. 1986. The teaching of learning strategies. In M. C. Wittrock (Ed.) *Handbook of research on teaching* (3rd Ed.) Macmillan, New York, NY, 315-327.

[9] Zimmerman, B. J., and Schunk, D. H. (Eds.). 2011. *Handbook of self-regulation of learning and performance*. Taylor & Francis.

[10] Siemens, G. 2005. Connectivism: A learning theory for the digital age. *Int. journal of inst. technology & distance learning* 2.1, 3-10.

[11] Van Hentenryck, P., Coffrin, C. 2014. Teaching Creative Problem Solving in a MOOC.*In Proceedings of The 45th ACM Technical Symposium on Computer Science Education.* ACM, New York, NY, 677-682.