Student Data: Data is knowledge – putting the knowledge back in the students' hands

Owen Corrigan, Dr Mark Glynn*, Aisling McKenna, Prof Alan Smeaton, Dr Sinead Smyth
Dublin City University, Ireland
**corresponding author*

Abstract
Learning Management Systems are integral technologies within higher education institutions. These tools automatically amass large amounts of log data relating to student activities. The field of learning analytics uses data from learning management systems (LMSs) and student information systems to track student progress and predict future performance in order to enhance learning environments (Siemens, 2011). The aim of this paper is to describe a project where we utilized a system developed in Dublin City University to use information about student engagement with our LMS, Moodle, to create a model predicting pass or failure in certain modules.

The project is divided into three distinct phases. An initial investigation was completed analyzing Moodle activity for the last six years. The purpose of this exercise was to determine automatically if "trends" could be identified linking Moodle engagement with student attainment. This was done by training a machine learning classifier to map student online behaviour, against outcomes. Once the classifier was trained, several modules were identified as suitable for building a predictor of student exam success.Ten modules were identified for semester 1 with a further seven identified for semester 2. The second phase involved analyzing current students' engagement with these modules and sending students information about the predictions of their attainment for the module, based on their Moodle engagement.

At this stage concerns were raised within the university that the data that we share with the students could actually have the opposite effect to what we are after, i.e. the student may look at the data and think that there is no point in putting in more effort as 'I'm too far behind already'. Dietz-Uhler and Hurn refer to this as "instead of being a constructive tool, feedback becomes a prophet of failure" (Dietz-Uhler, 2013). This contention was addressed by conducting an online survey with students in an effort to explore their experiences of being provided with feedback regarding their engagement with the LMS.

The third and final phase of this project was the development of a dashboard for lecturers to enable monitoring of their students' engagement with their module on Moodle. This enables lecturers to have an overview of how students are engaging with their course on Moodle and quickly identify students who are not engaging with the LMS and who are potentially at risk of failure or non-completion.

There are numerous examples of the use of learning analytics in higher education. This study focuses on the provision of data obtained through learning analytics to the student and qualitative analysis that was conducted in relation to this data. This research adds to the existing research into learning analytics being used for student retention.

Keywords: *Learner, student-retention, learning analytics, data-mining, VLE*

Introduction

The term digital footprint is a term that has emerged in recent years, particularly as the use of technology has become so ubiquitous in everyday life. Every "post" people make, every click they take, can be tracked and mapped back to an individual. As technology advances and specifically as the field of data mining advances, it has become easier to link data from a wide range of data sources back to a unique identifier, therefore increasing the power of each set of data points, essentially personifying the fact that the value of the entire dataset is greater than "the sum of the parts". In the field of data analytics, the ideal scenario is to link all data points back to one unique identifier, and in learning analytics this typically is a person's student number. Student data such as exam performance, attendance, socio-economic data and LMS engagement data can be combined not only to give a summary of a student's performance but with appropriate algorithms can be used to predict a student's performance in the future.

LMS's are integral technologies within higher education institutions. These tools automatically amass large amounts of log data relating to student activities. The field of learning analytics can use data from LMS's and student information systems to track student progress and predict future performance in order to enhance learning environments (Siemens, 2011). Analysing this data to determine if any

correlation exists, then summarising it and presenting it in a useful format is what is known as data mining. The more data points that you have at your disposal to "mine" the more accurate the aforementioned prediction of future performance will be. The advances of online learning and the move towards "blended" learning in the typical full time on campus courses have inevitably created more data that can be tracked, therefore making the predictions more accurate. So even though the concept of data mining is not new, the realisation and acceptance of its value has increased dramatically in recent years particularly in higher education.

In the field of technology enhanced learning the authors strongly believe that educators should lead with the pedagogy not the technology. In the case of learning analytics the authors believe that the researchers should lead with the learner/learning and not with the data. By extension in other words, learning analytics should now combine (or at least should consider the influence of) pedagogy and psychology and not just data mining. The psychological aspect of learning analytics refers to the impact that this data can have on a person's attitude. The challenge with "returning" data to students is the unknown effect that this data will have on the student going forward. Our research team consisting of a professor in data analytics, a data analytics researcher, a pshycology lecturer, the universitys' institutional research and data analysis officer and the head of teaching enhancement investigate the potential application of learning analytics within our university. This paper analyses the impact that the data generated through this project has on a student's attitude and performance

Research question
What is the impact of presenting students with data on their engagement with the college VLE?

Methodology
The data from the previous six years of the VLE was captured and student grades from their final exams were mapped to the VLE engagement data over the six year period to determine if there was some correlation between VLE engagement and exam performance. Where correlation was confirmed, algorithms were generated to use this data as a predictive tool to predict a current student's performance in their final exam based on their engagement with the VLE. For each module, we divided the Moodle logs into chunks of weeks. We then extracted multiple features for each week:

- A count of how many times they accessed Moodle
- The ratio of on campus to off campus accesses using IP address
- The average time of day they used Moodle
- How many times times they accessed Moodle during the weekend

These features have been shown to be good indicators of student performance (Casey et al., 2010). Work by Cocea and Weibelzahl on feature extraction was an influence on our approach (Weibelzahl & Cocea, 2006). The features used in this study include number of pages accessed, time spent reading pages and student performance on mid-semester tests. Even though they were using a VLE targeted at one particular module, the features used were similar to ours. However their objective was different, with their goal being to estimate learners' levels of motivations.
We used a Support Vector Machine (SVM) to classify students as a pass or a fail. We trained one SVM for each week of the semester up until the exams. Each training set contained all data available up until that week. For example the "week 7" SVM was trained with the demographic data, and all weekly Moodle log data up until week 7. This allows us to make predictions on new Moodle log data on a week by week basis. The following process was used:

- We selected only the first week of Moodle log data,
- We used 10-fold stratified cross validation
- We scored our model using Receiver Operating Characteristic Area Under Curve (ROC AUC) metric. This was used because it is less biased than accuracy when using imbalanced classes.
- We repeated this process for every week, including all of the Moodle data up to that week. We then graphed the performance of the classifier over each week.

This process gave us some intuition on whether the classifier was effective. If the performance of the classifier improved over time then the classifier was effective. We also noted that good classifiers typically achieve an ROC AUC score of above 0.6. We used these heuristics to select courses to target with intervention strategies for the coming semester. When an SVM makes a prediction for a

particular student it also produces a confidence score for that individual student. We can take advantage of this to rank the students according to their perceived need of an intervention.

Once the modules were identified and the algorithms were optimised for each model we then started to provide data to students on a weekly basis via email. At the end of the semester we conducted a quantitative analysis to determine if the weekly emails had an effect on a student's final grade for the module in question. To further analyse the effect of presenting the data to students we conducted a survey of students to gauge their opinion of the personal impact of receiving this data. The following sections outline the findings of these elements of the research obtained to date and outline plans for future research.

Limitations/Delimitations

Only five years of historic LMS data was used. As the project evolves over time, more data will be available. The algorithms will be continually revised therefore increasing the accuracy of the predictors of student performance. Furthermore, as this is a pilot study, the sample size is restricted to a small subset of students. While the study involves in excess of 1200 students we acknowledge that the quality of the study will be improved as we increase the number of students involved in this project. In addition the authors acknowledge that not every click is equal in terms of a student's engagement with a module. For example a 'click" to download a document from a course may not represent the same level of engagement with a course if a student responds to or initiates a discussion forum posting. Finally to strengthen the validity of the results of the online survey conducted with students, we plan to conduct a series of focus group interviews with students in an effort to triangulate the data from this research.

Results & Discussion

Figure 1 displays the total Moodle activity from 2010 up to and including 2013. The repeatability of the activity is very clear to see and the periodicity can be easily mapped to key times within the semester e.g. mid-semester exams, reading week, final exams etc. This repeatability indicated that it may be feasible to build a tool that may be able to connect performance in end of semester exams with students' VLE engagement. Moodle activity at a modular level was also analysed to see if there was similar evidence of periodicity and repeatability. As expected this was not visible for all modules, as a variety of factors beyond our control may have changed over the time in question e.g. different lecturers delivering the module, change in course content and/or structure etc.

However as illustrated in figure 2, numerous modules did exhibit the desired characteristics. The intensity changes visible between the different years can be explained by the difference in student numbers from one year to the next, in addition to the one class potentially being more "active" than another. Numerous modules were identified, based on the heuristics mentioned above. Nevertheless as this was a pilot study we decided to concentrate our efforts on a subset of these modules. We determined the subset on the basis of the pass rate at the end of the semester for the module and the year of study. We chose modules with a pass rate of less than 0.95 as we wanted to determine if there was a positive effect on the end of semester grades. In other words we wanted modules where there is room for improvement in the students' grades. The final criteria was to choose 1st year students, as retention of first year students is a high level strategic target for the university (Moore-Cherry et al., 2015).

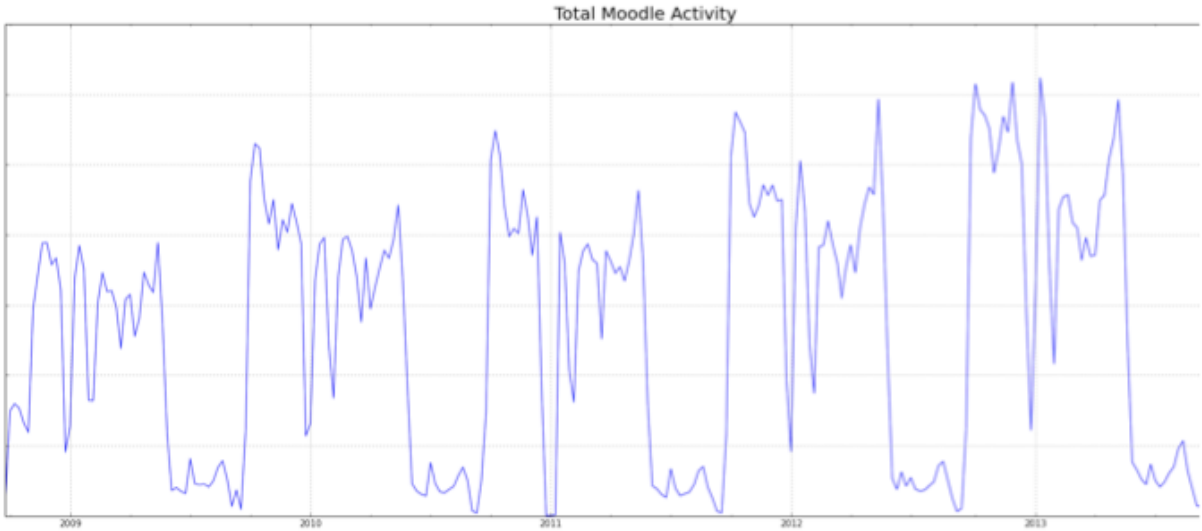

Figure 1 represents the total activity on the entire Moodle from 2009 up to and including 2013.

Figure 2 represents the total Moodle activity for one particular module (module D)

One of the challenges that we faced was having a control group against which we could compare our test group. We addressed this and concerns from the university ethics group by offering the students the option to opt in or out of the study by answering a simple poll available through the VLE before gaining access to the course content. Participation in the poll is mandatory but it was made very clear to students both in writing and verbally that there would be no difference in terms of access to course content whether they opted in or out of the study. Therefore our control group (n=377) were the students who opted out and our test group (n=1184) were those students who opted in. Once the control group was identified it was necessary to ensure that the two groups were comparable in terms of the academic ability of each group. In other words as we are comparing the final grades of students in particular modules we had to make sure that the stronger students, academically speaking, were not just in one of the groups. To assess this we measured the students' performance in the state examination for mathematics (Leaving Certificate Maths) for both groups. We chose this value as there is a very strong link between prior academic achievement in Mathematics and successful progression to the second year of higher education (Mooney et al., 2010). As illustrated in Figure 3 there is no significant difference between the students in both groups.

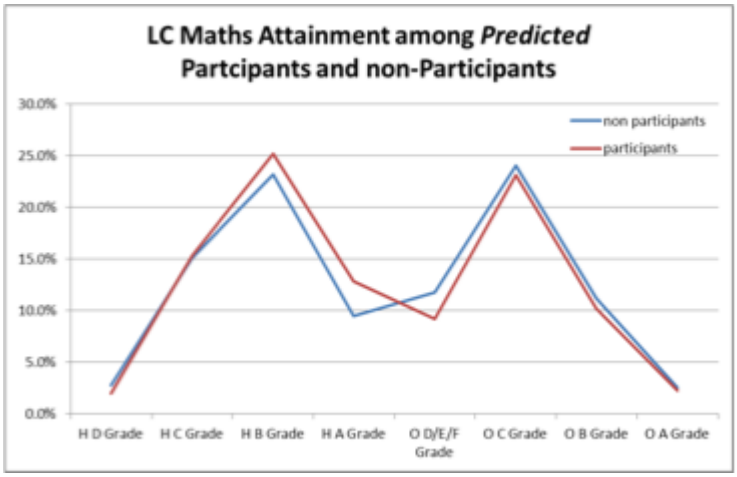

Figure 3: Leaving cert maths grades for participants and non-participants in the study

Once the validity of the control and the test groups were confirmed, the next task was to analyse the predictors generated by the algorithms, specifically looking at how many weeks the semester progressed before the predictions were deemed reliable. As mentioned above, the reliability of the predictors is very significant once the ROC value is above 0.6. In the case of the data represented in Figure 4 it is clearly visible that after four weeks the predictors are deemed to be reliable; in other words predictions can be made as early as week four with regards to a student's chances of success in that module. This is a very significant finding as currently our first real indication of a student at risk of non-completion is their first set of continuous assessment results which is normally half way through the semester but can be as late as week eight or nine. In this case we are finding out in half the time. This earlier notification allows us to "intervene" earlier and hopefully address the issues much more quickly to increase the student's chances of success completing the module. Tinto stated that the first six weeks are crucial in the experience of 1st year students in higher education (Tinto, 1988). Therefore to have a tool that can make a prediction and hopefully support students as early as the fourth week would be of great benefit.

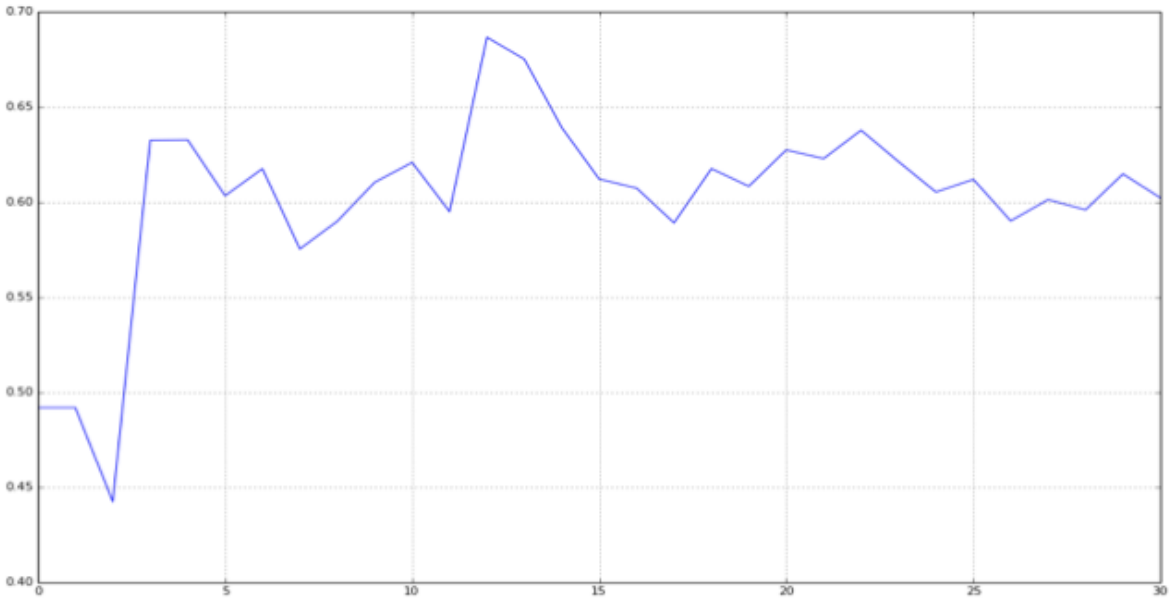

Figure 4: AUC ROC curve for module "H"

It should be noted that not all courses provide accurate predictors as early as week four; in some cases it was week six and on one occasion as late as week twelve. Therefore our next step was to look at the course structure and design on the VLE and the types of activities and resources that are available through the VLE.

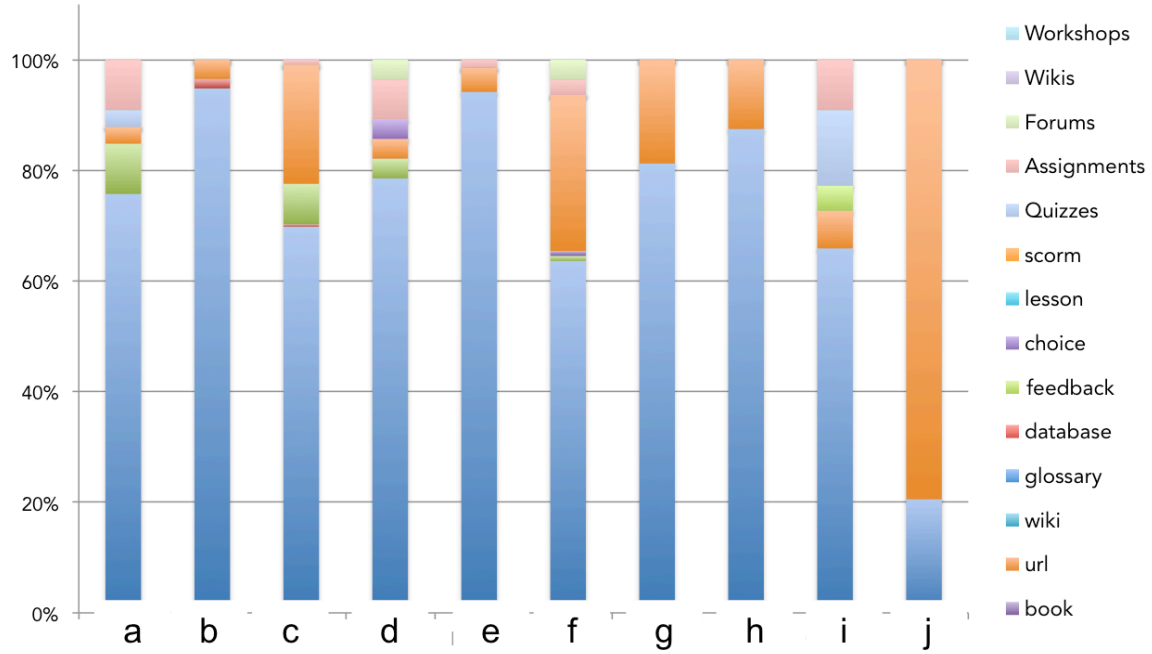

Figure 5 Summary of the module content available for each module on the VLE

Figure 5 illustrates that the majority of the modules analysed for this study are used mainly as document repositories and use very little of the VLE functionality. The exception is module "i" which contained a small selection of online quizzes for students to complete. While it is not desirable for the VLE just to be used primarily as a document repository, for the purpose of this study it is valuable to have this consistency across the modules. As each of the aforementioned modules on the VLE are essentially the same it does imply that the *way* the VLE is being used, i.e. different features of the VLE, is not a factor. The two remaining elements that can lead to variation in the accuracy of the predictors is: *when* the VLE content is released to the students, and the *volume of resources* that are placed on the VLE and their relevance to the module assessment. The former element is an unknown variable at this point but will be taken into account during the next stage of this project. The latter will be discussed after we discuss the overall result of this project, namely the impact of our intervention on the student's final grades for the various modules. To determine this we looked at the difference between those who opted-in to receiving emails each week, compared to those who opted-out. Table 1 shows this comparison and the last row indicates that for those who opt-in they can expect to see an average increase of +2.67% in their actual exam marks, all other things being equal. The final column of Table 1 shows the impact of PredictED on a per-module basis with ⁂ indicating significantly higher performance for participants (students) and ⁎ indicating higher performance. This table shows that for three modules, B, D and G, there was a significant improvement in students' exam performance while for five other modules there was improved performance. Only 47 students (of a possible 67) registered to take part in PredictED as part of F. This low number of registrations for the module as a whole was unexpected but we decided to continue and make the weekly alerts available nonetheless. The average decrease of 0.5% in exam performance can be discounted based on the low sample number for this group.

Table 1: Performance in end-of-module examination showing number of eligible students for module (Regs), and comparing average non-participants' exam mark (Non) vs. participants' average exam mark (Part.) and significance of difference between these (Signif.)

Code	Regs	Non	Part.	Difference	Signif.
A	154	45.70%	44.90%	-0.8%	-
B	157	60.80%	69.40%	8.6%	⁂
C	172	53.30%	54.90%	1.6%	⁎

D	152	59.40%	63.30%	3.9%	⁂
E	288	60.60%	61.80%	1.2%	⁎
F	47	67.00%	66.50%	-0.5%	-
G	284	58.90%	62.10%	3.2%	⁂
H	104	55.30%	57.00%	1.7%	⁎
I	122	70.30%	71.30%	1.0%	⁎
J	78	63.80%	65.30%	1.5%	⁎
Across all modules		58.58%	61.25%	2.77%	⁂

As mentioned earlier, another variable between the modules is the volume of content available to the students. Is there enough content for them to interact with, and does more content mean a more significant difference in the grades achieved by the students? Table 2 illustrates that there is no correlation between the number of resources / activities available on the VLE and the final grades difference between non-participants and participants. It is also worth noting that there is no correlation to the academic discipline of the module and the difference between the final grades of the participants and non-participants, although this observation would be strengthened if the number of modules studied was increased.

Table 2: Number of resources/activities available for students and the difference in grade between participants and non-participants within this study

Code	No. of resources / activities	Difference	Signif.		Code	No. of resources / activities	Difference	Signif.
H	24	1.7%	⁎		G	64	3.2%	⁂
D	28	3.9%	⁂		E	69	1.2%	⁎
A	33	-0.8%	-		J	73	1.5%	⁎
I	44	1.0%	⁎		F	110	-0.5%	-
B	58	8.6%	⁂		C	232	1.6%	⁎

The final element of this research at this stage was to contact the students involved to gauge their opinion on the impact of receiving the weekly notifications associated with this project. Students who took part were asked to complete a short survey at the start of Semester 2 - N=133 (11% response rate). The results of this survey are summarised in the table below

Over 70% who responded would take part again or they have been offered and are taking part again for the second semester

33% said they changed how they used the VLE. They claimed to have studied more, some comments include:

- "I read more articles online",
- "it proved useful for getting tutorial work done"
- *"I tried harder to engage with my modules on loop"*

- *"I think as it is recorded I did not hesitate to go on loop. And loop as become my first support of study."*
- *"I logged on more"*

Finally most of the students said that they also used the VLE more for modules outside of the study. This is an unexpected positive outcome. However this raises the issue of unexpected outcomes and what is known as the Hawthorne effect. The Hawthorne Effect is a term associated with unexpected outcomes from a study which are believed to depend on the fact that the subjects in a study have been aware that they are part of an experiment and it influences their behaviour (Merret, 2006). In our study this means that the students change their use of the VLE because they know that they are being watched, not necessarily because of the notifications that we send them. As the Hawthorne Effect can be a result of subconscious decisions in addition to conscious decisions the effect cannot be isolated for the purpose of this study. Nevertheless the purpose of our study is to increase students engagement with the VLE in the hope that this will result in an increase in their final grades for their modules. Therefore it is not our concern whether the increase in engagement is due to the notification that we send students or if it is the Hawthorne Effect – the end result is the same. It is worth noting that if the reason for increase in engagement is due to the Hawthorne Effect there is a possibility that this will fade over time. However as our study is limited to first years and essentially we get a new "audience" every year the potential impact of a fading hawthorne effect is not signficant. To measure the influence of the notification that we send students we created two different types of notifications. The results of the impact of this variation will be discussed in a future paper.

Conclusions & Recommendations

The main goal of the initiative was to improve student retention by sending weekly notifications to students to give them an indication of their engagement with the VLE and encouraging them (when appropriate) to engage more with the VLE. This goal was achieved because there was an average increase in grades of 2.9% when comparing the grades of participants in this study with grades of non-participants across the various modules.

The students were very positive in their opinion about the initiative and over 70% will participate in the study for other modules should the opportunity arise.

The significance of the Hawthorne effect could not be disentangled from the other variables associated with this study. Its significance will only become clearer when a longitudinal study of the data involving students as they progress throughout their degree is conducted (as opposed to just first year as is the case in this report).

The majority of the modules on the VLE examined through this study are designed in the same way, i.e. as document repositories. Therefore it is feasible to now look at the modules in more detail to determine if we can idenitfy factors related to module design that influence how early in the semester our predictors become accurate.

Acknowledgements

This research was supported by Science Foundation Ireland under grant number SFI/12/RC/2289, and by Dublin City University.

Bibliography

Casey, K., Gibson, P. & Paris, I.-S., 2010. Mining Moodle to understand Student Behaviour. In NUIM, ed. *In International Conference on Engaging Pedagogy 2010 (ICEP10)*. Maynooth, 2010. ICEP.

Dietz-Uhler, B.&.H.J., 2013. Using Learning Analytics to Predict (and Improve) Student Success: A Faculty Perspective. *Journal of Interactive Online Learning*, 12(1).

Merret, F., 2006. Reflections on the Hawthorne Effect. *Journal of Educational Psychology*, 26(1).

Mooney, O., Patterson, V., O'Connor, M. & Chantler, A., 2010. *A Study of Progression in Irish Higher Education*. Dublin: Dept of Education HEA.

Moore-Cherry, N., Burroughs, E. & Quinn, S., 2015. *Why Students Leave: Findings from Qualitative Research into Student Non-Completion in Higher Education in Ireland*. HEA.

Siemens, G., 2011. *February 27–March 1, 2011, <>(2011)*. [Online] Available at: https://tekri.athabascau.ca/analytics/ [Accessed 10 May 2015].

Tinto, T., 1988. Stages of Student Departure: Reflections on the Longitudinal Character of Student" . *The Journal of Higher Education*, 59(4), pp.438-55.

Weibelzahl, S. & Cocea, M., 2006. Can log files analysis estimate learners' level of motivation? In *Lernen - Wissensentdeckung - Adaptivitat (LWA2006),*. Hildesheim, 2006. University of Hildesheim.

Woosley, S.A., 2003. How Important Are the First Few Weeks of College? the Long Term Effects of Initial College Experiences. *College Student Journal*, 37(2).