A CASE STUDY OF LEARNING ANALYTICS WITHIN A STATISTICS COURSE FOR UNDERGRADUATE STUDENTS IN ECONOMICS

Catherine Dehon, Philippe Emplit and Emma Van Lierde
Université libre de Bruxelles
cdehon@ulb.ac.be

Higher education institutions globally face a continuous expansion of their enrolment in which learner success constitutes a major challenge. Therefore, there is growing interest in the analysis of data linked to student learning engagement. Indeed, large amounts of learning-related student data are currently not being fully exploited, while their aggregation and quantitative analysis would definitely be elements valuable to support teachers and students, to optimize students' learning experience. In this global context, we have applied, in a public university without any academic filter for enrolment, such analysis to virtually tutor first-year undergraduate students in a statistics course. By supporting them in the form of voluntary online self-assessing tests, we examined what were the personal profiles of the students who were using available tests and how they exploited this help. Finally, using econometric models we tried to determine if there was a link between student success and the use of this help.

INTRODUCTION

Belgium is a federal state where the language-based communities (French-, Dutch- and German-speaking) are competent for the educational system. The French-speaking community (FWB) offers a unique framework for the analysis of success at university. The main highlighted features of this system are the openness of enrolment at any university degree to (almost) all socioeconomic and educational backgrounds, the non-existent grade publishing requirements and very low, common tuition fees (an important part of higher education is financed through public funds). As a result, almost 70 percent of the student population that finishes the general high school system enrols at university. However, during the first year, very high rates of failure and drop-out are observed (Arias & Dehon, 2013). Moreover, statistical courses are present in many curricula, including in the humanities and social sciences, where the skills and prerequisites in mathematics as well as student motivation for basic courses in statistics are extremely variable. It is therefore very important for this type of course to be able to best help students to success.

In this specific community and institutional context, we would like to assess, by the way of learning analytics, how some virtual help, especially designed for early-stage learners by their educators and posted online on the institutional learning management system (LMS), the university's educational platform, is used by students and whether or not the use of these tools has an impact on their success. Indeed, there has been since 2011 an increasing interest in learning analytics (LA), i.e. the acquisition and the quantitative analysis of data linked to learners' academic activity, enabling higher education institutions to exploit large amounts of student data that were previously not used to their full potential, with among other sources the Learning Management System (LMS) (Leitner, Kahlil, & Ebner, 2017; Siemens & Long, 2011). This aggregation and analysis of these data enables the support of institutions' main stakeholders, namely learners, instructors and the administrative staff, for improving student experience and eases the understanding of the current situation and actions to be taken to achieve such improvement (Siemens & Long, 2011). Examples of learning analytics use has been seen worldwide, with tools such as "Ma Réussite" at the Université Laval in Québec enabling all stakeholders, i.e. learners, educators or academic officials, to take appropriate actions, on the basis of some student's performance coloured indicators. In this case, those are related to the student individual use of online resources compared to the aggregation of the use of his/her peers, and they are generally considered as adequate predictors of final grades (Pothier, 2016). A similar tool was used at Purdue University and seems to have yielded positive results, with higher retention and higher performance for students who used the tool (Sclater, Peasgood, & Mullan, 2016).

By using a unique data set containing the entire enrolled undergraduate student population in economics at the *Université libre de Bruxelles* (ULB), this case study aims to be the first complete analysis to investigate which online elements might be considered, be useful and be consistently developed in the context of the creation of a predictive model for student success, based on the use by

In: S Budgett (Ed.), *Decision Making Based on Data*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
August 2019, Kuala Lumpur, Malaysia. iase-web.org/Conference_Proceedings.php

the learner of the institutional educational platform (LMS) in terms of frequency and intensity (Van Lierde, 2018). Therefore, the research questions are threefold. Firstly, it questions whether there are any common characteristics among the students who use the available offered help, and if so, what those characteristics are. Secondly, it looks at how these students use the available help, namely with which frequency, intensity and consistency, with the attempt of inferring their purpose. Thirdly, it analyses whether the use of the available help impacts student success at the exam.

DATA AND METHODOLOGY

The case study is based on undergraduate students in economics officially registered at the *Université libre de Bruxelles* during the academic year 2017-2018 and having in their annual study program the first course of statistics (638 students). Introduced online help for this course included one prerequisites test and four course chapter tests which were not mandatory to be fulfilled by the student. The prerequisites test was introduced on the LMS. Moreover, four course content tests were proposed on the LMS as well, each of which to be taken on a voluntary basis and could lead to a 0.5 bonus point (on a maximum of 20 points) on the final grade of the course in case of success with more than 70%. This set of tests was introduced in hopes that students would study regularly along the academic year and better succeed in the final exam.

The data set for each student comes from two main sources, namely the student information system and the learning management system. The former gives access to student personal data, such as gender, year of birth, nationality, scholarship status, former high school and registration information, including grades, among others. From the last source was extracted information about student's online activity within the aforementioned tests but also regarding his/her overall engagement in terms of time spent on the course material and in terms of login frequency. To find the determinants that influence the student's behaviour with respect to the use of available help for the course of statistics and its impact on student's success, we use descriptive statistics and econometric models taking into account individual characteristics, prior schooling and socioeconomic background.

EMPIRICAL RESULTS

The first research question is focused on the profile (personal, academic and socioeconomic characteristics) of the students using the help put at their disposal. The subgroups are defined based on how effectively a student uses the help, namely "no help", "some help" and "all help". By using the available help, we mean that the student tried one or more times to fulfil the offered tests, including the prerequisites and the content tests. Among the 638 registered students, 99 used none of the available help, 266 used some help and 273 used all the offered help.

Table 1. Some descriptive statistics use of available help

Variables	Modalities	Frequency	No help	Some help	All help
Population 2017-18	All students	638	16%	42%	43%
Gender	Male	402	20%	45%	36%
	Female	236	8%	37%	55%
Years repeated in high school	On time	187	6%	33%	61%
	1 or more years "late"	451	20%	45%	35%
Newly enrolled student	Yes	354	9%	33%	58%
	No	284	23%	53%	24%
High school diploma	CESS general	405	14%	37%	49%
	CESS technic	68	21%	51%	28%
	Foreign diploma	145	15%	49%	36%
High school type	FWB no discrimination	350	12%	38%	51%
	FWB discrimination	115	23%	43%	34%
	Outside FWB	153	15%	50%	35%

The variables used (Table 1) are based on individual characteristics (gender, newly enrolled student in higher education), prior schooling (years repeated in high school and high school type of

diploma) and socioeconomic background (socio-economic level of the high school). The CESS (*Certificat d'enseignement secondaire supérieur*) diploma delivered by high schools in FWB could be classified into "general" and "technic". The first is the usual degree to enter at university, while the second type of diploma is more oriented towards technical professions. Being in Brussels, ULB also attracts a large proportion of students with foreign secondary education diploma. Concerning the type of high school, we classified the schools into three types, namely those within the FWB with no positive discrimination, those within the FWB with positive discrimination and those outside of the FWB. The positive discrimination is given to schools welcoming students coming from poorer socioeconomic environments. Schools in this category receive additional means from the government. As results, we find that male students, "late" students, students who have already failed the course of statistics, students with vocational/technic high school background and students coming from schools with positive discrimination, use significantly less help than those coming from "regular" schools in the FWB. In conclusion, we observe that the students who might really need help, those who are late at university, who have retaken the class, or who have a vocational high school degree, are the ones who do not use it intensively.

To further deepen the analysis of the potential impact of student personal characteristics and previous academic track record on the extent to which they use the offered help, we run an ordinal logistic regression with three levels for the dependent variable: "no help", "some help" or "all help" (see Table 2). After a procedure of selection, we analyse the effect of the following variables on the use of available help by students: their gender, whether they are Belgian or not (linked to the high school location), their time of arrival at university, whether they are "first generation" or not, and what type of high school they went to.

Table 2. Ordered logistic regression on the use of available help

Ordered logistic regression (Stata software)					**95% Conf. Interval**	
Variable	**Odds Ratio**	**Std. Error**	**z**	**p-value**	**Min**	**Max**
Female	**2,224**	0,377	4,72	**0,000**	1,596	3,099
Belgian	**0,651**	0,136	-2,06	**0,040**	-0,432	0,980
Late	0,763	0,166	-1,24	0,215	0,497	1,170
Newly enrolled student	**3,507**	0,687	6,41	**0,000**	2,389	5,147
FWB discrimination	**0,446**	0,097	-3,71	**0,000**	0,290	0,683
Foreign diploma	**0,462**	0,106	-3,38	**0,001**	0,296	0,724
/cut1	-1,860	0,339			-2,524	-1,195
/cut 2	0,463	0,331			-1,863	1,113
LR chi2(6)	122,18					
Prob > chi2	0,000					
Pseudo-R^2	0,098					

At the level of the model, the likelihood ratio test testing the assumption of proportional odds computes a p-value of 0.3951 confirming the fact that this model is in accordance with this necessary assumption. The model shows a pseudo-R^2 of 9.8%, which is quite low but common in research in education: this leads us to the conclusion that other important variables could be pertinent to explain the participation such as for example the motivation. Nevertheless, student's past education and individual characteristics of student introduced in the model partly explain the use of help. Being a woman or a "first generation" student are characteristics that increase the probability of using more help, while being from a high school with positive discrimination or outside of the FWB decreases the probability of using more help. Being Belgian is significant in the model, but it is linked to the school variable and should thus be interpreted cautiously: here, it looks like it has a negative impact. Being "late" at university is not significant in the model.

For the second research question, we dig deeper into the analysis of the use that students make of the available help, including the used attempts, the time spent on using the help and how well they succeeded the tests. For those tests where a bonus is possible, we try to assess whether that seems to

be a driver behind the use of the help. We also look at the overall use of the LMS for the course in terms of login frequencies and total time spent on the course's webpage. We note that only few students who managed to get their bonus in the first attempt used the test a second time leading us again to believe that the main student objective is to receive the bonus point, and not to improve their score afterwards. Surprisingly, some of the students who did not get the bonus in the first attempt did not try the test a second time. Of those who have not attempted the exam, two thirds had failed to gain their bonus point. We finally look at the overall use of the LMS by students in terms of login frequency and duration on the platform for the course. We look at four different variables for all students since we consider the term period and the exam period. On average, students spent about 8.5 hours on the course page during the term, logging in 17 times, and about 3 hours during the exam period logging in 4 times. However, differences across students are very large.

To analyse the third research question, the impact of the use of the available help on student success at the final evaluation of the course we first use descriptive statistics of the student's exam performance within the different groups defined by the level of help they use. Of the 129 students who passed the first attempt exam (20% of success), 100 used all available help. In other terms, nearly 50% of the students who used all the offered help succeeded instead of less than 8% for the other students (students who used none of the available help or some help). However, we keep the potential selection bias in mind, since the characteristics of the group using the help are not the same of those of the group not using the help. Indeed, as in many studies in the social sciences, it is impossible to set up a randomized experiment where a group of randomly selected students would be obliged to use the help (group treated), and the other students could not have access to these aids (control group). To investigate this issue, we use the Heckman model (Amemiya, 1984), which is a sample selection model. This type of model is used when the dependent variable is known only for a part of the sample (truncated sample), sub-sample selected according to a previous choice and not randomly. This model assumes that there are two decisions, one being to select oneself into the group, which is a discrete decision (using all/some help or not), and the other being the result of the studied outcome continuous variable (the exam score). We do not work with such datasets in this case, but we have students who choose to use the help and those who do not, and we would like to assess whether there is a selection bias as well as the effect of variables on the exam score. If we suppose we did not know the exam score for students who did not use the available help, we could manually create a truncated dataset in which we only observe the exam score for the students who have used the help to assess whether there is a selection bias that we should account for if choosing to perform a regression on the exam performance of students who used some or all help only. If the bias is insignificant, it would suggest that we have controlled for all variables impacting both the choice of the extent to which the student uses the help and the performance. Then classical linear regression model could be performed directly on data, the selection bias being insignificant.

In the Heckman model (Table 3), there are 91 "artificially" censored observations, i.e. students who did not use any of the available help under the condition that we had information about their previous high school, and 527 uncensored observations. The model shows that there is no significant selection bias with using the traditional level of significance of 5% (Wald test (rho=0): chi2 = 3,13; p-value = 0.0771). Confirming for what we had seen in the first part, significant variables for the use of help include the gender, whether the student is first generation and whether he/ she was in a positively discriminated high school. When using these variables, the Heckman model suggests that errors terms of the two equations (participation equation and exam scores equation) are uncorrelated leading to the conclusion of no selection bias. That suggests that we have not omitted any important variables that would impact both the choice of the extent to which the student uses the help and the performance.

As no major selection bias exist, we perform linear regressions, considering the exam score as dependent variable and the student characteristics, use of help and the use of the online platform as independent variables (Table 4). Four models are estimated with different control variables.

Table 3. Heckman model output

Heckman model output (Stata software)						
Exam score					**95% Conf. Interval**	
Variable	**Coef.**	**Robust Std. Error**	**z**	**p-value**	**Min**	**Max**
Female	**0,518**	**0,334**	**1,550**	**0,121**	**-0,137**	**1,174**
Late	**-1,058**	**0,438**	**-2,410**	**0,016**	**-1,916**	**-0,199**
Newly enrolled student	**-0,110**	**0,392**	**-0,030**	**0,978**	**-0,780**	**0,758**
FWB discrimination	**-0,611**	**0,453**	**-1,350**	**0,177**	**-1,499**	**0,276**
Foreign diploma	**-0,371**	**0,392**	**-0,950**	**0,344**	**-1,139**	**0,397**
Constant	**5,746**	**0,558**	**10,300**	**0,000**	**4,652**	**6,840**
Use help					**95% Conf. Interval**	
Variable	**Coef.**	**Robust Std. Error**	**z**	**p-value**	**Min**	**Max**
Female	**0,444**	**0,144**	**3,090**	**0,002**	**0,162**	**0,727**
Late	**-0,283**	**0,201**	**-1,410**	**0,159**	**-0,677**	**0,111**
Newly enrolled student	**0,495**	**0,155**	**3,190**	**0,001**	**0,191**	**0,799**
FWB discrimination	**0,457**	**0,160**	**-2,850**	**0,004**	**-0,772**	**-0,143**
Foreign diploma	**-0,136**	**0,157**	**-0,870**	**0,387**	**-0,443**	**0,171**
Constant	**1,030**	**0,230**	**4,470**	**0,000**	**0,579**	**1,182**
rho	**-0,296**	**0,157**			**-0,567**	**0,330**
Censored obs	**91**		**Wald test**	**chi2(1) = 3,13**		
Uncensored obs	**527**		**(rho=0)**	**p-value = 0,077**		

Table 4. Linear regression models for student success

Exam score (n=618)	**Model 1**		**Model 2**		**Model 3**		**Model 4**	
Variable	**Coef.**	**p-value**	**Coef.**	**p-value**	**Coef.**	**p-value**	**Coef.**	**p-value**
Female	**0,87**	**0,005**	0,16	0,564	-0,04	0,865	-0,11	0,637
Belgian			**0,73**	**0,027**	0,53	0,089	0,36	0,204
Late	**-1,29**	**0,001**	**-1,01**	**0,003**	**-0,97**	**0,002**	**-0,99**	**0,001**
Newly enrolled student	0,3	0,397	**-0,97**	**0,003**	**-0,99**	**0,001**	**-1,04**	**0,000**
FWB discrimination	**-1,07**	**0,008**	-0,26	0,466	-0,15	0,647	-0,09	0,766
Foreign diploma	-0,45	0,210	0,49	0,179	**0,77**	**0,027**	**0,79**	**0,013**
Some help			1,93	**0,000**	-0,43	0,349	-0,59	0,166
All help			5,44	**0,000**	**1,39**	**0,022**	0,36	0,540
Average score					**0,53**	**0,000**	**0,29**	**0,000**
LMS intensity year							**1,01**	**0,000**
LMS intensity exam							**0,29**	**0,046**
Constant	4,94	0,000	1,65	0,006	1,71	0,003	3,62	0,000
Adj R-squared		**0,0585**		**0,3057**		**0,3815**		**0,4826**

In the first model, only student characteristics are used as control leading to a quality of the prediction of the exam score rather small (R^2 = 0.0585) although most of the explanatory variables are significant. Model 2 suggests that if a student goes from “no help” to “some help”, the exam score is expected to increase by 2 points in average (exactly 1,93), and from “no help” to “all help”, the exam score is expected to increase by 5,44 points in average (the maximal score for the exam being 20). When accounting for student performance, here computed as the average score of the four content

tests, the model explains 38% of the variability in the data and past characteristics have smaller impacts. The use of help is still significant, although only when using all the help available and with a smaller impact on the expected exam score. This variable reflects probably the motivation and the student's engagement in his/her studies, while the average score will rather reveal the student's skills. So, it seems like a student who has a little more difficult and thus a lower average score on the tests is able to compensate part of that difficulty by the engagement in using all the available help. When adding the use of the LMS during the year and during the exam period, the use of help becomes insignificant, which might be because the use of help is linked to the use of the LMS. A higher use of the LMS overall has a positive impact on the performance at the exam. The last model is able to explain up to almost 50% of the variability of the exam score.

DISCUSSION

When looking at how the students use the available help for the course of statistics posted on the LMS, we find that those who use the self-assessing tests perform well on average. Unfortunately, we observe that the students who might really need help, those who are late at university, who have retaken the class, or who have a vocational degree, are the ones who do not use it intensively. Yet the use of help can somewhat diminish the initial "drawbacks" linked to student past characteristics. But it must be remembered that evaluating the impact of aids to success is always a complicated exercise because of possible selection bias. In this context, we have used Heckman model to take into account such difficulties. Logically, we also find that student term performance has a big impact in the prediction of future performance. The estimates from the models are obviously dependent on the context and the type of student but it is reasonable to think that in other situations, the same tendencies could be identified. In addition, the proposed methodology can be directly applied in other contexts.

To conclude, and despite the difficulties related to the diversity of teaching methods, legal issues and computer difficulties of automation of these systems, we definitely support recommendations such as aggregating databases and setting up learning data-based dashboards for students and for teachers. A more systematic use of learning analytics could really help students in their academic experience.

REFERENCES

Amemiya, T. (1984), Tobit Models: A Survey. *Journal of Econometrics*, *24*, 3–61.

Arias, E., & Dehon, C. (2013). Roads to success in the Belgian French community's higher education system: Predictors of dropout and degree completion at the Université Libre de Bruxelles. *Research in Higher Education*, *54*(6), 693–723.

Leitner, P., Khalil, M., & Ebner, M. (2017). Learning Analytics in Higher Education - A Literature Review. In A. Peña-Ayala (Ed.), *Learning Analytics: Fundaments, Applications, and Trends,* (pp. 189–211). New York, NY: Springer International Publishing.

Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE Review*, *46*(5), 30.

Pothier, F. (2016). Appui à la réussite des étudiants à l'Université Laval - Des outils facilitant le dépistage préventif des étudiants en situation de difficulté.

Sclater, N., Peasgood, A., & Mullan, J. (2016). Learning analytics in higher education: A review of UK and international practice: Case Study A: Traffic lights and interventions: Signals at Purdue University.

Van Lierde, E. (2018). On students' use of Learning Management System tools in higher education: A case study of learning analytics within the ULB. *Master thesis*. Brussels.