What Can Be Found from Student Interaction Logs of Online Courses Offered in Brazil

André L. B. Damasceno, Cássio F. P. Almeida, William P. D. Fernandes, Hélio C. V. Lopes, Simone D. J. Barbosa

[1] Pontifícia Universidade Católica do Rio de Janeiro – Rio de Janeiro, RJ – Brasil

{adamasceno,calmeida,wfernandes,lopes,simone}@inf.puc-rio.br

Abstract. *Online Education has broadened the avenues of research on student behavior and performance. In this paper, we compare results in the literature about student behavior patterns and performance with an analysis of VLE logs of online courses offered in Brazil. We conducted a study exploring and analyzing data using statistical methods and machine learning techniques on a dataset provided by a Brazilian institution. Then, we compared the results of our analysis with what the literature says about student behavior and performance. Finally, we show that, although most results related to student access and course completion can also be found in courses offered in Brazil, some of our results contradict existing work, mostly when related to student performance.*

1. Introduction

Distance Learning (DL) is not a novelty. For instance, in Brazil, there are records of correspondence courses in the 19th century [Saraiva 1996]. Years later, with the rising of computers and evolution of the Internet, instructors and students have experienced new ways of teaching and learning. One of the reasons for the increase in supply and demand for online courses is that students can determine their own study pace and participate in courses regardless of geographic distance limitations [Seaton et al. 2014].

Students' interactions with Virtual Learning Environments (VLEs) are often stored in logs. The analysis of these logs can predict the students' performance, evaluate their learning achievement in a course, and even identify behavior patterns [Romero and Ventura 2010, Dutt et al. 2017]. We conducted a systematic mapping [Damasceno et al. 2019] to compile the main problems, objectives, methods, case studies, and results presented in papers that discuss the use of logs to analyze and predict both student behavior and performance. However, all results presented in the mapping were from courses offered outside Brazil.

The purpose of this paper is to identify which results in the literature can be found in online courses offered in Brazil. To achieve our goal, we have explored and analyzed, using statistical methods and machine learning techniques, a dataset provided by a Brazilian institution that offers large-scale online courses. In general, results found in the literature related to student access and course completion can also be found in courses offered in Brazil. However, most analyses about student performance did not show results in line with existing work.

This paper is structured as follows. Section 2 shows an overview of the results found in the literature used in this research about student interaction logs analysis. Section 3 describes both the courses and dataset used, as well as the methods adopted in the

DOI: 10.5753/cbie.sbie.2019.1641

data analysis. Section 4 presents the procedure adopted to analyze the student interaction logs and discusses the results of the analysis. Lastly, Section 5 presents some final considerations.

2. Related Work

Logs from VLEs are a gold mine to understand students' behavior and performance. The exploration of these logs has recently leveraged much research on Informatics in Education. For instance, some papers have shown that students can be clustered into different groups based on their access or interaction patterns, such as, with which resources they interact first, how long they stay or at what times they usually access the VLE [Guo and Reinecke 2014, Park et al. 2017, Khosravi and Cooper 2017]. In addition, some research has shown that access to the online environment resources increases in periods close to exams or assignment deadlines [Nandi et al. 2011, Park et al. 2017, Shi et al. 2017]. Complementing that, Chen and Zhang claimed that student inactive for more than continuous 3 weeks, likely had drop out from the course [Chen and Zhang 2017].

In regard to students' interactions in forum, Nandi et al. showed an increase in forum posts close to deadlines [Nandi et al. 2011]. Other results have shown that: (i) student groups that use more forums tend to have a good performance [Nandi et al. 2011, Carter et al. 2017]; (ii) student groups that have more posts are more likely to complete the course [Andres et al. 2018]; (iii) student groups that complete more assignments tend to use more forums [Kizilcec et al. 2013]; (iv) forum usage can be used as a predictor of students completing the course [Andres et al. 2018]; (v) in general, women post more than men [Crues et al. 2018]; and (vi) student groups that initiate threads in forums tend to complete the course [Andres et al. 2018] and have a good performance [Carter et al. 2017].

Lastly, some research has presented classification models to predict, through the analysis of students' interaction, their performance [Carter et al. 2017, Shi et al. 2017, Al-Shabandar et al. 2018]. Some researchers have found that page viewing times are significantly correlated with students' final score [Zhang and Zhu 2017] and drop out [Shi et al. 2017, Laveti et al. 2017]. Others stated that successful students are more frequently engaged with online assignments and participate regularly [Guo and Reinecke 2014, Shi et al. 2017, Al-Shabandar et al. 2018, Boulton et al. 2018], assignment completion is a cue that the student will achieve good performance, and it can also be used as a predictor of course completion [Andres et al. 2018, Al-Shabandar et al. 2018]. For instance, Al-Shabandar et al. presented positive correlations between productive, assignment completion, and pass rates [Al-Shabandar et al. 2018]. Finally, Laveti et al. and Al-Shabandar et al. showed that data access can be used as a predictor of students completing the course [Laveti et al. 2017, Al-Shabandar et al. 2018].

3. Dataset and Methods

Our dataset was provided from UNASUS-UFMA[1], an institution that offers large-scale online courses, mostly in the Healthcare area. They include undergraduate, specialization, and training-on-demand courses. All the courses use the Moodle[2] software platform. The

[1] `http://www.unasus.ufma.br/`
[2] `https://moodle.org/`

dataset comprises three file types: i) *logs file*, which provides a timestamped log of every student's interaction with the system, *e.g.*, viewing course materials, interacting with the forum, or any other activity; ii) *grades file*, which provides both the students' grades and gender; and iii) *schedule file*, describing the course schedule, including the task deadlines. All user data were anonymized and the students are identified by unique ids in both *logs* and *grades files*.

Table 1 shows an overview of the dataset, consisting of 5 specialization courses (*Saúde da Pessoa Idosa*, *Saúde da Família*, *Atenção Básica*, *Nefrologia* and *Atenção Domiciliar*) offered between 2013 and 2017, with a total of 755,869 records. In particular, three courses were offered more than once (*Saúde da Pessoa Idosa*, *Saúde da Família*, *Atenção Básica*). Each course is organized in two or three cycles and all cycles are composed of a set of modules. In general, these modules are led by instructors (*i.e.*, teachers and tutors), who provide the content in ebooks (available to download in PDF format) and some modules (1 or 2 per course) make use of video and audio resources. Instructors evaluate the students through their postings related to topic discussions in the forum, assignments submitted and quizzes answered. Such evaluations result in numerical values used as part of the module grade. In the end of each cycle, students take a test in a physical classroom, and this grade is used to calculate the final grade of the modules included in the cycle. As a requirement to conclude the course, the students have to achieve module grades greater than or equal to 7 and write a final paper, whose presentation is also in a physical classroom. Apart from the test and final presentation in the classroom, all other course activities are online.

Table 1. Overview of the dataset.

Course	Edition	Cycle	Modules	Period	Students	Log entries
Saúde da Pessoa Idosa	1	2	23	11/2013 - 01/2015	206	61,632
	2	2	23	07/2014 - 09/2015	253	74,842
Saúde da Família	1	3	15	12/2013 - 01/2015	179	27,442
	2	3	15	04/2014 - 02/2015	200	30,618
Atenção Básica	1	3	15	03/2014 - 06/2015	279	42,696
	2	3	15	07/2014 - 08/2015	224	34,275
	3	2	15	09/2014 - 08/2015	332	50,800
	4	2	15	12/2014 - 12/2015	839	128,372
	5	2	15	04/2015 - 02/2016	402	61,512
	6	2	15	04/2015 - 02/2016	112	17,139
	7	2	15	10/2015 - 09/2016	146	28,541
	8	2	15	05/2016 - 07/2017	335	57,289
Nefrologia	1	3	12	10/2014 - 03/2016	454	77,642
Atenção Domiciliar	1	2	13	04/2015 - 07/2016	289	63,069

In order to verify whether the literature results presented in Section 2 could be reproduced with our dataset, we first conducted an exploratory data analysis to understand, extract, and organize the meaningful data. For instance, the dataset provides 95 user interaction types (*e.g.*, course view, assignment submission, posting on forum). However, only

64 have records related to student interactions. The others were recorded by instructors' interaction, whose analysis is outside the scope of this paper. Therefore, only 64 interaction types were used as features in our analysis. Moreover, we noticed that there are no access records to resources included by the teacher (*e.g.*, ebooks, videos and audios), because all of them were accessed through links to other pages outside of the VLE (*e.g.*, YouTube, ebook repository system).

Then, we counted the records of each feature by student, arranging them per course, edition, and cycle. To improve our analyses, we derived features to identify: i) *students who completed the module*, based on the grades file which do not present grades of students who dropped out; ii) *students who completed the course*, verifying who completed all modules and has final paper grades; iii) *students' results in each module*, identifying whether the student has a final module grade greater than or equal to 7; iv) *students' results in course*, checking whether the student has all final module grades and final paper grade greater than or equal to 7; v) *number of days each student accessed the modules*, based on students' access during the cycle period; vi) *number of days accessed until the in-class test day*, based on students' access during the cycle period; and vii) *student inactivity for three or more continuous weeks in the course*. We also built a dataset with the number of student accesses by day, arranging them per course and edition.

As all features are represented by categorical and numerical data, statistical methods (*e.g.,* Pearson correlation, Wilcoxon rank test) were used in their analysis and interpretation. The statistical analysis makes inferences about each result presented in the literature on student interactions.

Finally, we also aimed to identify student clusters based on the interaction patterns and develop models to predict both students' performance and drop out. Traditional statistical analyses develop accurate prediction models based on human input in making assumptions about the relationships between variables. Therefore, we used machine learning techniques due to their capabilities to analyze high dimensional log data, of arbitrary form, characterized by both noise and complex non-linear pattern components. The choice of techniques such as K-means, Logistic Regression, Random Forest and Decision Tree used in the development of clustering and prediction models were based on techniques widely used in the literature [Damasceno et al. 2019]. In addition, we evaluate our prediction models using as quality measure Precision (Prec.), Recall (Rec.), and F1-Measure (F1).

4. Analysis and Results

In this section, we present the analysis procedure and discuss the results. We took into consideration the students': i) access and interaction patterns, ii) forum usage, iii) performance, and iv) course completion. In general, we correlated access, interaction patterns and forum usage with the Moodle features that recorded student interactions. We define as student performance their result in the in-class test and course completion, if they presented the final paper. Moreover, the analyses used the dataset grouped by cycle (total of 33) and course edition (total of 14).

First, we used a K-means clustering algorithm to analyze whether the students could be clustered based on their interaction logs in the VLE. The elbow method was used to detect the number of student interaction clusters per cycle through computing and plotting the sum of square errors in order to identify where the marginal gain drops significantly, producing an angle (elbow) in the graph. In line with existing work [Guo and Reinecke 2014, Park et al. 2017, Khosravi and Cooper 2017], we found 3 or 4 clusters (depending on the course, edition, and cycle). We then noticed that, in all clusters, the feature that records course access (*i.e.,* course view) is the most meaningful to discriminate the clusters. However, there are no significant differences of student performance across the clusters. In other words, the means of test grades are similar in all of them. For instance, Figure 1 shows the dispersion of course access and test grades by cluster in the *Saúde da Família* course, edition 1, cycle 1. Complementing that, we removed from the dataset the no-show students in the in-class test and, using Pearson correlation, we did not find correlation between the number of page viewing (*i.e.,* course view) and the test grades, contradicting the result found by Zhang and Zhu (2017).

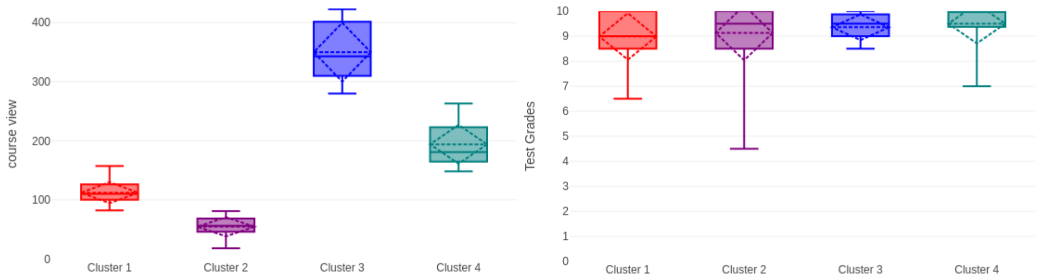

Figure 1. Dispersion of course view and test grades by cluster in *Saúde da Família* course, edition 1 and cycle 1.

We noticed that the number of accesses increases in periods close to in-class tests and assignment deadlines, in line with existing work [Nandi et al. 2011, Park et al. 2017, Shi et al. 2017]. In particular, for in-class tests, we analyzed a period of 15 days before each test date, whereas for the assignments, we analyzed a period since the date the assignment was available until its deadline. To do this, we used the generalized linear mixed-effects model (GLMM) [Bates et al. 2015] and we fit it to the dataset with the number of student accesses by day. Our analysis used the slope to identify the access trend and we defined as increasing trend the positive slopes with significance level of 5%. As described in the Table 2, 67% of the periods preceding assignment deadlines and 82% of periods close to the in-class tests had an increasing trend. In general, the number of accesses increased on average 1.13 and 1.07 per day before the assignments and in-class tests, respectively. Besides, as presented in Table 3(i), we found in only three courses inactive students for three or more consecutive weeks (*i.e*, *Saúde da Pessoa Idosa* edition 1 and 2, *Atenção Básica* edition 2). Most inactive students did not complete the course, so inactivity can be used as a cue of course drop out, in line with existing work [Chen and Zhang 2017].

In regard to forum interactions, we analyzed whether the use of forums was related to students' performance. It is worth noting that forum access is required for all forum activities (*e.g.*, add a post, view comments). Therefore, to this analysis, we correlated the students' forum access (*i.e.*, forum view discussion) and test grades. In addition, we removed the no-show students in the in-class test from the dataset, grouped by cycle. We did not find a significant correlation in most courses, contradicting existing work [Nandi et al. 2011, Carter et al. 2017]. We used Pearson correlation and we found a significant

Table 2. Trends analysis of access in the VLE in periods preceding in-class tests or assignment deadlines.

	Number of periods in which accesses...		
Periods close to	**increased**	**decreased**	**showed no trend**
assignment deadlines	**289 (67%)**	50 (12%)	91 (21%)
in-class tests	**27 (82%)**	1 (03%)	5 (15%)

correlation ($p < .0001$) only in the *Saúde da Pessoa Idosa* course, albeit a low correlation, ranging from .32 to .46, depending on the cycle. Then, using the dataset grouped by course edition, we analyzed whether students who have more posts are more likely to complete the course [Andres et al. 2018]. We split the students by course completion, and through the Wilcoxon rank test and Pearson correlation, we noticed a strong correlation in all courses ($p < .0001$). However, in all courses, the forum was used in assignments and as part of the module grade and the students cannot add forum topics. We therefore did not analyze whether students who completed more assignments tended to use forums more [Kizilcec et al. 2013], whether there was correlation between initiating threads and course completion [Andres et al. 2018], and whether the number of forum posts rose in periods close to deadlines [Nandi et al. 2011].

Table 3. Analysis of (i) inactive students for continuous three weeks or more, (ii) correlation between number of accesses and student success (passed/failed).

		(i)		(ii)		
Course	**Edition**	**Dropped out**	**Concluded**	**Passed**	**Failed**	**p**
Saúde da Pessoa Idosa	1	**72**	48	124	82	**< .0001**
	2	**22**	17	130	123	**< .0001**
Saúde da Família	1	0	0	153	26	**< .0001**
	2	0	0	177	23	**.0242**
Atenção Básica	1	0	0	222	57	**< .0001**
	2	**1**	0	154	70	**< .0001**
	3	0	0	158	174	**< .0001**
	4	0	0	638	201	**< .0001**
	5	0	0	199	203	**< .0001**
	6	0	0	71	41	**.0002**
	7	0	0	91	55	**.0005**
	8	0	0	252	83	**.0005**
Nefrologia	1	0	0	321	133	**< .0001**
Atenção Domiciliar	1	0	0	123	166	**.0093**

Table 4 presents the results of the analysis whether women have more postings than men using the dataset grouped by course edition, keeping in mind that in most courses there were more women than men. Using the Wilcoxon rank test and Pearson correlation, we have not found a significant difference of postings according to gender: in some courses, on average men posted more than women, and in others the opposite occurred. This result contradicts existing work [Crues et al. 2018]. Using the same dataset and statistical methods, we analyzed whether there was a correlation between successful students (*i.e.*, who passed in the course) and number of accesses. As shown in Table 3(ii), our results showed that successful students had more page views than failing students, in line with existing literature [Guo and Reinecke 2014, Shi et al. 2017, Al-Shabandar et al. 2018, Boulton et al. 2018]. Next, we verified whether there was correlation between students' success and completion rate of online assignments. To do this, we used the dataset grouped by cycle without no-show students in the in-class test and split the students by whether they passed or failed. We applied the same statistical methods taking into account the students' grades and assignment features (*i.e.*, quiz attempts, assignment submissions, assignment submissions for grading, forum posts). As a result, we noticed that only 6 of the 33 cycles presented a significant correlation. In most courses, there was no positive correlation between online assignments and pass rates, contradicting existing work [Al-Shabandar et al. 2018].

Table 4. Analysis whether women had more postings than men.

		Number of		**Mean of posts**		
Course	**Edition**	**Women**	**Men**	**Women**	**Men**	**p**
Saúde da Pessoa Idosa	1	**174**	32	**22.218**	17.781	.162
	2	**209**	44	15.377	**17.318**	.085
Saúde da Família	1	**111**	68	17.468	**17.602**	.866
	2	**119**	81	**16.042**	14.617	**.011**
Atenção Básica	1	**168**	111	**20.041**	16.891	**.013**
	2	**113**	111	**22.283**	19.630	.160
	3	**209**	123	**13.598**	12.934	.123
	4	**553**	286	**18.488**	16.045	**.001**
	5	197	**205**	**16.446**	12.829	**< .001**
	6	**61**	51	**17.032**	15.196	.150
	7	**81**	65	15.296	**16.046**	.705
	8	**202**	133	**17.519**	15.909	**.016**
Nefrologia	1	**374**	80	**16.540**	16.037	.831
Atenção Domiciliar	1	**234**	55	11.931	**12.000**	.916

We also built classification models to predict both students' course completion and performance. In general, the dataset used was grouped by course, and by course and cycle. However, we grouped the *Atenção Básica* course into 5 distinct groups, because it had been offered in 5 distinct ways: (i) the editions 1 and 2 had 3 cycles, whereas the others had 2; (ii) the editions 3 and 4 were designed with 6.5 modules in the first cycle and 8.5 in the second cycle; (iii) the editions 5 and 6 had 4.5 modules in the cycle 1, 7.5 in the cycle 2, and 3 modules out of cycles; (iv) the edition 7 had the same number of modules (total of 7.5) in both cycles; and (v) in edition 8, the cycles comprised 8.5 and 6.5 modules respectively. We split our data per course into 80% to train and 20% to test.

As mentioned in Section 3, we created three models for the prediction tasks: Logistic Regression, Random Forest, and Decision Tree. In order to calibrate our models, we used a 5-fold cross-validation over the training set. Table 5 presents the quality of the best model for the proposed prediction tasks. Each column presents the best model (on average) for each task.

Table 5. Mean of the best model results to predict course completion using: (i) all features, (ii) features related to assignment submissions and quiz attempts, (iii) data accesses and (iv) features related to forum. (v) Results of the prediction model of students' performance using all features.

	(i)	(ii)	(iii)	(iv)	(v)
Method	Random Forest	Decision Tree	Decision Tree	Decision Tree	Decision Tree
Prec.(%)	84.58	81.93	83.18	74.65	37.23
Rec.(%)	79.76	78.33	78.63	71.72	35.13
F1(%)	79.34	78.08	78.28	70.43	32.72

First, we built models to **predict students completing the course**, using the dataset grouped by course with all features. Our Random Forest classifier had $F_{\beta=1} = 79.34\%$ (Table 5(i)), in line with some papers that presented accurate classification models to predict student drop out rates through their interaction in the VLE [Shi et al. 2017, Laveti et al. 2017]. Next, using only features related to assignment submissions and quiz attempts, we analyzed **whether assignment submissions could be used as a predictor of course completion** [Andres et al. 2018, Al-Shabandar et al. 2018]. As presented in Table 5(ii), the Decision Tree had $F_{\beta=1} = 78.08\%$, only 1.26% less than the model that used all features (Table 5(i)). We also built models using only features related to **data access (*i.e.*, course views) to predict course completion** [Laveti et al. 2017, Al-Shabandar et al. 2018]. As we see in Table 5(iii), the best performing model was the Decision Tree, with $F_{\beta=1} = 78.28\%$, 0.2% greater than the model using only features related to assignment submissions and quiz attempts Table 5(ii), and 1.06% less than the model that uses all features (Table 5(i)). Then, we used all **forum features to predict course completion** [Andres et al. 2018]. Table 5(iv) shows the best model, Decision Tree, with $F_{\beta=1} = 70.43\%$, a decrease of 8.91% compared to models that use all features (Table 5(i)), 7.65% compared to models that use features related to assignment submissions and quiz attempts Table 5(ii), and 7.85% compared to models that use features related to data access (Table 5(iii)). That shows that using only forum features had a negative impact on model quality. Lastly, we built models to **predict student performance through their interaction in the VLE** [Carter et al. 2017, Shi et al. 2017, Al-Shabandar et al. 2018]. To do this, we used the dataset grouped by course and cycle. We segmented the grades into four groups: (i) lower than 7, (ii) between 7 and 7.9, (iii) between 8 and 8.9, and (iv) between 9 and 10. As we see in Table 5(v), $F_{\beta=1}$ was strongly impacted, decreasing to 32.72%. However, we could not directly compare this model to the other models, since the prediction objective was different: this model aimed to predict student final score, whilst the other ones aimed to predict student course completion.

5. Final Considerations

This paper compared the results found in the literature about the use of VLE logs to identify student behavior patterns and performance with those found in online courses offered in Brazil in the Healthcare area. We used data exploration, statistical methods and machine learning techniques.

Our analyses showed evidences that the students can be clustered by their accesses and successful students have more page viewing than failed students. We also found results showing that the number of accesses increased in periods close to exams and assignment deadlines. In addition, students who were inactive for three or more consecutive weeks can be used as a cue of course drop out. Besides, we noticed that students who had more posts were more likely to complete the course. In regard to the prediction models, we found good results related to course completion. Conversely, we did not find significant differences of student performance across the clusters. In regard to forum interactions, we also did not find correlations with student performance, and significant difference in postings across gender. Another result was that in most courses there was no positive correlation between online assignment submissions and pass rates. Lastly, the models built to predict the student performance did not achieve reasonable results.

As future work, we aim to identify student study styles through methods of Process Mining and verify whether there is a correlation between those styles and student performance. We also noticed a gap in regard to analyzing instructor behavior in VLEs. In future work, we plan to develop a dashboard using visual analytics techniques. To evaluate the dashboard, we will assess whether there are changes in student performance when instructors are able to see information about student behavior and performance, and act accordingly.

References

[Al-Shabandar et al. 2018] Al-Shabandar, R., Hussain, A. J., Liatsis, P., and Keight, R. (2018). Analyzing Learners Behavior in MOOCs: An Examination of Performance and Motivation Using a Data-Driven Approach. *IEEE Access*, 6:73669–73685.

[Andres et al. 2018] Andres, J. M. L., Baker, R. S., Gašević, D., Siemens, G., Crossley, S. A., and Joksimović, S. (2018). Studying MOOC completion at scale using the MOOC replication framework. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, LAK '18, pages 71–78, New York, NY, USA. ACM.

[Bates et al. 2015] Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1).

[Boulton et al. 2018] Boulton, C. A., Kent, C., and Williams, H. T. (2018). Virtual learning environment engagement and learning outcomes at a 'bricks-and-mortar' university. *Computers & Education*, 126:129–142.

[Carter et al. 2017] Carter, A. S., Hundhausen, C. D., and Adesope, O. (2017). Blending measures of programming and social behavior into predictive models of student achievement in early computing courses. *ACM Trans. Comput. Educ.*, 17(3):12:1–12:20.

[Chen and Zhang 2017] Chen, Y. and Zhang, M. (2017). Mooc student dropout: Pattern and prevention. In *Proceedings of the ACM Turing 50th Celebration Conference*, ACM TUR-C '17, pages 4:1–4:6, New York, NY, USA. ACM.

[Crues et al. 2018] Crues, R. W., Henricks, G. M., Perry, M., Bhat, S., Anderson, C. J., Shaik, N., and Angrave, L. (2018). How do Gender, Learning Goals, and Forum Participation Predict Persistence in a Computer Science MOOC? *ACM Transactions on Computing Education*, 18(4):1–14.

[Damasceno et al. 2019] Damasceno, A. L. B., Ribeiro, D. S., and Barbosa, S. D. J. (2019). What the literature and instructors say about the analysis of student interaction logs on virtual learning environments. In *2019 IEEE Frontiers in Education Conference (FIE)*.

[Dutt et al. 2017] Dutt, A., Ismail, M. A., and Herawan, T. (2017). A Systematic Review on Educational Data Mining. *IEEE Access*, 5:15991–16005.

[Guo and Reinecke 2014] Guo, P. J. and Reinecke, K. (2014). Demographic differences in how students navigate through moocs. In *Proceedings of the First ACM Conference on Learning @ Scale Conference*, L@S '14, pages 21–30, New York, NY, USA. ACM.

[Khosravi and Cooper 2017] Khosravi, H. and Cooper, K. M. (2017). Using learning analytics to investigate patterns of performance and engagement in large classes. In *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education*, SIGCSE '17, pages 309–314, New York, NY, USA. ACM.

[Kizilcec et al. 2013] Kizilcec, R. F., Piech, C., and Schneider, E. (2013). Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. In *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, LAK '13, pages 170–179, New York, NY, USA. ACM.

[Laveti et al. 2017] Laveti, R. N., Kuppili, S., Ch, J., Pal, S. N., and Babu, N. S. C. (2017). Implementation of learning analytics framework for MOOCs using state-of-the-art in-memory computing. In *2017 5th National Conference on E-Learning & E-Learning Technologies (ELELTECH)*, pages 1–6. IEEE.

[Nandi et al. 2011] Nandi, D., Hamilton, M., and Haland, J. (2011). How active are students in online discussion forums? In *Proceedings of the 30th Australasian Computing Education Conference - Volume 114*, ACE '11, pages 125–134, Darlinghurst, Australia, Australia. Australian Computer Society, Inc.

[Park et al. 2017] Park, J., Denaro, K., Rodriguez, F., Smyth, P., and Warschauer, M. (2017). Detecting changes in student behavior from clickstream data. In *Proceedings of the 7th International on Learning Analytics and Knowledge*, LAK '17, pages 21–30, New York, NY, USA. ACM.

[Romero and Ventura 2010] Romero, C. and Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618.

[Saraiva 1996] Saraiva, T. (1996). Educação a distância no brasil: lições da história. *Em Aberto*, 16(70):17–27.

[Seaton et al. 2014] Seaton, D. T., Bergner, Y., Chuang, I., Mitros, P., and Pritchard, D. E. (2014). Who does what in a massive open online course? *Commun. ACM*, 57(4):58–65.

[Shi et al. 2017] Shi, Y., Peng, Z., and Wang, H. (2017). Modeling student learning styles in moocs. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, pages 979–988, New York, NY, USA. ACM.

[Zhang and Zhu 2017] Zhang, M. and Zhu, J. (2017). A data-driven analysis of student efforts and improvements on a spoc experiment. In *Proceedings of the ACM Turing 50th Celebration Conf.*, ACM TUR-C '17, pages 1:1–1:6, New York, NY, USA. ACM.