Data-Driven Analysis of Engagement in Gamified Learning Environments: a methodology for real-time measurement of MOOCs

Khulood Alharbi[1*], Laila Alrajhi[1] and Alexandra I. Cristea[1], Ig Ibert Bittencourt[2], Seiji Isotani[3] and Annie James[4]

[*]Corresponding author, khulood.o.alharbi@durham.ac.uk.
[1] Computer Science, Durham University, Durham, UK.
[2] Federal University of Alagoas, Alagoas, Brazil.
[3] University of Sao Paulo, Sao Paulo, Brazil.
[4] University of Warwick, Warwick, UK.

Abstract. Welfare and economic development is directly dependent on the availability of highly skilled and educated individuals in society. In the UK, higher education is accessed by a large percentage of high school graduates (50% in 2017). Still, in Brazil, a limited number of pupils leaving high schools continue their education (up to 20%). Initial pioneering efforts of universities and companies to support pupils from underprivileged backgrounds, to be able to succeed in being accepted by universities include personalised learning solutions. However, initial findings show that typical distance learning problems occur with the pupil population: isolation, demotivation, and lack of engagement. Thus, researchers and companies proposed gamification. However, *gamification design* is traditionally exclusively based on theory-driven approaches and usually ignore the data itself. This paper takes a different approach, presenting a *large-scale* study that analysed, statistically and via machine learning (deep and shallow), the first batch of students trained with a Brazilian gamified intelligent learning software (called CamaleOn), to establish, via a grassroots method based on learning analytics, *how gamification elements impact on student engagement*. The exercise results in a novel proposal for *real-time measurement on Massive Open Online Courses (MOOCs)*, potentially leading to iterative improvements of student support. It also specifically analyses the *engagement patterns* of an underserved community.

Keywords: Grassroots Method, Data-driven Approach, Gamification.

1 Introduction

Education is a major part of society [1] and a fundamental human right. Nevertheless, access to it, especially to higher education (HE), varies. For example, in the UK, higher education is widely available (around 50% in 2015-16); however, in the Brazilian context, every year millions of students compete to have the opportunity to study in a high-quality, public university (with only up to 20% succeeding). Although we believe that the long-term solution to this problem is to improve the quality of basic education for all, the current burden in balancing the discrepancy between public and private education is taken over in countries like Brazil, due to the distances involved, by organisations providing e-training for the subjects of the entrance exam for universities. However, most of the solutions follow the classical and inappropriate *one-size-fits-all* approach, which has been shown to be inadequate, and, especially in MOOCs, results in a massive dropout rate [2] [3].

Gamification [4] has been proposed as a potential solution. However, previous research proceeded by implementing these, based, at best, on theoretical considerations [5], and from the few that proceeded to evaluation, the latter's aim was to prove the validity of specific gamification [6]. We believe that, whilst a theoretical basis, especially rooted in pedagogy, is necessary, it is essential to ground the design process itself on lessons learned, in a cyclic manner, from the usage of the system and the learner behaviour. Thus, in this paper we propose to *redesign e-learning systems from the data itself, and let data guide the recommendation of new features*. Moreover, as large-scale studies are few and far between, this research offers an invaluable insight into the issues and opportunities inherent in scaling such systems.

This paper therefore presents a large-scale study that analysed the first batch of students trained with a Brazilian gamified intelligent learning software (CamaleOn[1]), to establish and describe how gamification elements and resources impact on student engagement, and inform thus a redesign of the system, based on grassroots educational data mining. From a technical point of view, the aim of this research is to *move from theory-driven to a data-driven approach, in order to redesign effective gamified intelligent learning systems*. In terms of application, this work aims at a vital problem of our society, that of *sustainable personalised inclusive large-scale distance learning for boosting the chances of disadvantaged groups to pass the challenging public tests for admission into prestigious universities in Brazil*.

Thus, the main research questions is:
Do gamification elements increase engagement in MOOCs? and if so, *how can we find out in real-time which gamification elements impact on student engagement?*

2 Related Research

2.1 Gamification

With the use of e-learning systems like MOOCs, initial findings show that typical distance learning problems occur with the pupil population, such as: isolation, demotivation, and lack of engagement [7]. For this reason, some researchers and companies propose the application of *gamification* to deal with such problems [8] [9] [10]. Gami-

[1] https://plataformacamaleon.com.br/

fication is defined as 'the use of game elements in a non-game context' [11]. Gamification has been widely disseminated not only in the context of research, but also in terms of business applications. However, the introduction of gamification in online learning is relatively new. Efforts are still being made to understand gamification and find the best use for game mechanics in learning environments [6]. In addition, whilst there are many known benefits about the use of gamification in education [12], the design of gamified learning systems is usually theory-driven. As a result, there is a *lack of runtime feedback*, *non-gamified scaffolding*, and *under-exploitation of interaction data*. Whilst the theoretical basis is very important in designing purpose-fit gamified systems, in the context of large-scale online learning, it is not feasible to propose a *one-size-fits-all* design of gamification. For this reason, it is very important to take into account the data generated from the system, in order to better understand the users' interactions, and refine the offering. [13] presented a taxonomy for gamification elements and their potential effect over student behaviour, like engagement and motivation, which was evaluated by experts' via surveys. In this paper we use a data-driven approach to analyse this effect using an online learning environment's log data. We analyse students' interactions with gamification elements and use machine learning classifiers to predict their effect over engagement, from early interactions, thus simulating a real-time analysis.

2.2 Engagement

The process of learning involves many elements contributing to its success, and one of these is *engagement*. Learner engagement is an important factor in academic performance. [14] define engagement as being incorporated in behavior, emotions and thinking. In both physical and virtual classrooms, learner engagement is a major factor of learner achievement. Literature has explored engagement and provided several approaches to improve it. Supporting and improving learner engagement has been shown to have a highly positive effect on academic performance [15]. [16] studied participants engagement in MOOC environment through analysing their reflective comments to understand what makes MOOC engaging, by applying five machine learning classifiers. Their findings highlight some practical solution for instructors that include strategies for supporting student-tutor interactions, such as motivating the students through enthusiastic attitude and using humour to arouse students' attention. In this research, the e-training course does not involve instructor interaction, thus we focus on other suitable and effective elements to improve and increase engagement.

2.3 Educational data mining and learner analytics

Educational data mining [17] and learner analytics [18] [3] allow for a grassroots view of actual interaction between students and e-learning systems. These areas have been growing in popularity recently [19]. They have been used for student modelling or student behaviour modelling, prediction of performance, increase in (self-) reflection and (self-) awareness, prediction of dropout, retention, improved assessment and feedback services, recommendation of resources [20], or scientific inquiry, personalisation, domain modelling, grouping, planning and scheduling, and parameter estimation [21]. However, they have not been used, to the best of our knowledge, for the cyclic (re-) design of increasingly adaptive and gamified e-learning systems. This means that the discovered relations and rules would, at best, be used based on somewhat rigid initial assumptions on the existent system, instead of *inspiring a completely new re-design, just based on lessons learnt from real-time data*. We also boast *tapping into the great potential of the developing world*, and its specific landscape of educational needs.

3 Methodology

3.1 Approach

To understand how to improve gamification, we have a completely different approach to related research. Instead of building a system from scratch, based on existing or expanding theories, we analyse user behaviour in a given system of e-learning, and base our improvement suggestions on the existing user behaviour. In our case, this system is CamaleOn (see next section). However, the beauty of this approach is that it can be applied to any system. This is also a more realistic approach, as many educational online systems are available and in use, and it is a costly and often problematic to change them completely. Instead, a more gradual approach to this change is proposed, based initially on available data, and subsequently informed by gamification theories.

3.2 CamaleOn

CamaleOn is a Brazilian Gamified Intelligent Tutoring System. Officially launched in 2012, its aim is to increase the accessibility of educational resources to Brazilian students. There is a particular focus on providing students from public schools the resources needed to attend a Brazilian university. To motivate the user to continue with the website, CamaleOn uses different aspects of gamification (e.g., elements such as experience points (XP), badges, etc. as methods for motivation). Figure 1 presents the design of CamaleOn's webpage, where points (XP) are displayed on the top of the screen at all times, to provide a visualisation of the student's advancement via a gamified progress bar. Trophies are greyed out until earned; each holding a label explaining how it can be earned. Additionally, a progress map at the bottom of the screenshot visualises the student progress through the subjects of the curriculum.

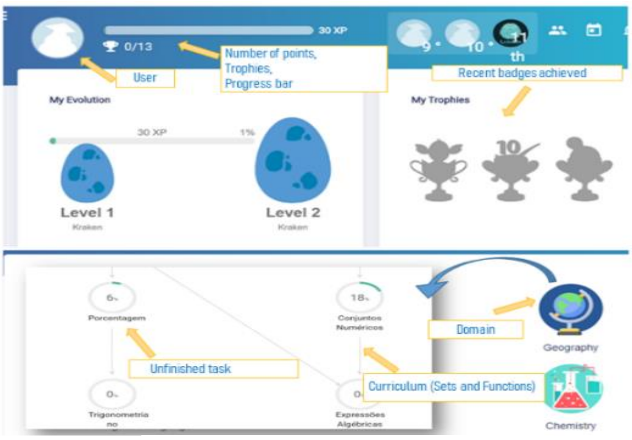

Fig. 1. CamaleOn: main pages.

3.3 Data

Data collected from CamaleOn represents 8270 students, a sample size much above the required statistically applicable one for the student population of Brazil (for confidence level 95%, confidence interval ±5%, sample size calculator from Surveysystem.com, for the population of Brazil at 211 million people[2] a minimum of 384 people is needed). Students solved 307814 problems, watched 1131 videos, received 236345 badges, logged in 67752 times. Data was collected on their behaviour (Logs) to build a Student Model [22]. Behaviour reflects interaction data between the students and the various elements of their online learning environment, such as problems, resources, etc.

3.4 Matching Data to Research Questions

The first step in the data-driven approach is to extract refined research questions from the data, based, on the overall aims of the research. In Table 1, the bold words in the "Data Subset" column indicate which dataset the subset of data originated from. The list of attributes, following the dataset, are the attributes which were selected from that subset. Analysing the attributes and data available from CamaleOn, we need to first extract the *gamification* elements used; here, they are badges, points, medals. For student *engagement*, we can use frequency of interaction (e.g., number of logins) and lack of dropout (thus involvement in the higher levels of the course).

Table 1. Matching data sets to research questions.

Data subset	Research items
Students: Number of Points, Number of Badges, Number of Medals, Number of Problems Solved, Number of Mistakes and Number of Correct Answers **Logs:** Log Type (equal to "Problem Solving"), Problem Correctly Done	Investigate performance of students versus engagement

The purpose is to find out if existent gamification features are useful, and if more gamification features need to be introduced, to address engagement. It is important to note here that further analysis is possible, and that this paper only illustrates how existent data may be used to improve the design of an extant system.

3.5 Definitions and Measures

For our research question, we chose to define engagement by both the number of logins and the total number of question attempts. Students' academic performance is not a necessary indication of engagement. Here, we set the threshold for the **highly engaged group** of students as consisting of students u $u \in St$ from the student cohort, where:

---

[2] https://www.ibge.gov.br/apps/populacao/projecao/index.html

$G_{HE}=\{u \in St | \#login_u \geq avg(\#login_{u; \forall u \in St})$ AND $\#questions_u \geq avg(\#questions_{u; \forall u \in St})\}$ (1)

Where *#x* refers to the number of x and *avg(y)* computes the average value of *y*.

This corresponds to students who have logged into the system more than 8 times and attempted to answer at least 304 questions, which are the mean values for number of logins and question attempts, respectively. This resulted in 1058 highly engaged students, and 7212 less engaged students. The gamification elements in the system are:

- **Points**: points are earned by answering low level questions.
- **Medals**: medals are earned by showing high skills in questions answering such as answering all questions in a topic correctly or solving side assignments.
- **Badges**: earned by interacting with the system in a specific way such as: spending one hour in the system or learning a sub-assignment 3 days in a row.

We define the gamification elements in the system by the variable "**Reward Count**" RC_u, as the sum of Points p_{qu}, Badges $b_{qu}$and Medals m_{qu}earned by a student *u*:

$$RC_u = \sum_{q=0}^{\#no\_que} p_{qu} + \sum_{q=0}^{\#no\_que} m_{qu} + \sum_{i=0}^{\#no\_int} b_{qu} \quad (2)$$

We first answered the research questions using correlation analysis, based on the Pearson coefficient. Next, we use both shallow and deep learning methods to further answer the questions in more depth. For shallow methods, we use and compare a number of ML models for classification: Linear Regression (LR), Linear Discriminant Analysis (LDA), K-Nearest Neighbours (KNN), Classification and Regression Trees (CART) and Naive Bayes (NB). Then we apply two deep learning algorithms to compare the performances of Machine Learning (ML) against Deep Learning (DL) models for numerical data with a low number of predictors, namely Multilayer Perceptron (MLP) and Convolutional Neural Network (CNN), which are recommended for numerical, non-sequential data. Figure 2 provide a general view to our methodology.

Fig. 2. General view of methodology followed in answering the research question.

4 Results

4.1 Normality Test

For the normality test of high and low engagement for students, we applied the Kolmogorov–Smirnov test, rather than Shapiro Wilk, due to the large data size that exceeded 5000 instances. Results indicate a non-normal distribution for each group (p ≤.00).

4.2 Data Visualisation: Higher/Lower Engagement versus Gamification Use

We next visualise the two groups, to analyse visual differences in gamification elements' use, via the total number of earned rewards for each group (Figure 3).

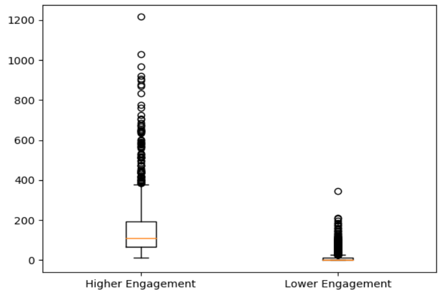

Fig. 3. Box plot higher and lower engagement groups versus earned rewards (points, badges, medals).

4.3 Data Correlation: Engagement versus Gamification

Table 2 shows the correlation between engagement and gamification. For instance, it indicates a strong positive association between students' number of logins and the number of rewards they earn. The highest correlation value is noticed between Badges and engagement status. The lowest value is seen between the number of earned medals and that of logins, possibly due to fact that medals are questions and curricula related. However, the engagement variable "Is Engaged" shows a positive association with all of the gamification elements represented by Reward Count, RC_u.

Table 2. Correlation test results between engagement indicators and gamification elements.

	Reward Count	Points	Medals	Badges
High login	0.531	0.482	0.373	0.631
High question attempts	0.660	0.656	0.604	0.671
Is engaged	0.660	0.656	0.604	**0.682**

4.4 Engagement Prediction based on Gamification

Following the correlation test results, we used the gamification elements and the additional aggregate parameter "Reward Count", and their combination, as inputs of different dimensions, to classify high and low engagement with various classification models (Table 3). The output of the classifier would either be the learner is engaged (1) or not engaged (0). These results show that the CamaleOn gamification elements are a strong predictor for students' engagement, with all accuracies > 0.924. I.e., the number of rewards students earn is strongly linked to the number of logins and general advancement through the system. The accuracy of CNN and MLP exceed the traditional ML models, suggesting that ML and DL classifiers perform slightly better - but similarly, for problems with a small number of features, such as this. MLP was the clear overall winner in terms of prediction model comparison. The highest score is observed (mostly) with the combination of all elements. What is interesting is the similarity of individual elements' score, despite the differences between them in functionality and purposes. I.e., Medals reward curricula advancement, while Badges reward defined system actions.

Table 3. Classifiers' results for engagement level based on gamification elements.

	Inputs	Acc	Low-engagement (0)			High-engagement (1)		
			P	R	F1	P	R	F1
LR	Reward Count	.951	.97	.98	.98	.87	.78	.82
	Points	.950	.97	.98	.98	.87	.77	.82
	Medals	.937	.95	.98	.97	.85	.65	.74
	Badges	.938	.96	.97	.97	.80	.72	.76
	All Elements	**.954**	.97	.98	.98	**.88**	**.80**	**.84**
LDA	Reward Count	.924	.93	**.99**	.96	**.98**	.46	.63
	Points	.950	.93	.98	.96	**.98**	.45	.62
	Medals	.937	.92	.98	.96	.96	.42	.59
	Badges	.938	.92	.97	**.97**	.80	**.72**	.76
	All Elements	**.954**	**.96**	**.99**	**.97**	.91	**.72**	**.80**
KNN	Reward Count	.947	.97	.97	.97	.81	.82	.81
	Points	.950	.97	.98	.97	.83	.78	.81
	Medals	.937	.96	.97	.96	.76	.75	.75
	Badges	.938	.96	.95	.96	.71	.75	.73
	All Elements	**.954**	**.98**	**.98**	**.98**	**.84**	**.84**	**.84**
CART	Reward Count	.944	.96	.98	.97	.81	.73	.77
	Points	.950	.96	.98	.97	**.83**	.69	.76
	Medals	.937	.95	.98	.97	.82	.67	.74
	Badges	.938	.96	.97	.97	.80	.72	.76
	All Elements	**.954**	**.97**	.97	.97	.81	**.81**	**.81**
NB	Reward Count	**.954**	.98	.97	**.98**	**.83**	.84	**.83**
	Points	.950	.98	.97	**.98**	**.83**	.69	.76
	Medals	.937	.96	.97	.97	.78	.74	.76
	Badges	.938	.96	.97	.97	.80	.72	.76
	All Elements	**.954**	.98	.96	.97	.78	.86	.82
MLP	Reward Count	.958	.98	.97	.98	.83	.84	.84
	Points	.957	.98	.97	.98	.83	.86	.83
	Medals	.942	.96	**.98**	.97	.82	**.71**	.76
	Badges	.941	.96	.97	.97	.80	.72	.76
	All Elements	**.964**	.98	**.98**	.98	**.87**	**.86**	**.86**
CNN	Reward Count	**.957**	.97	.98	.98	.85	.81	.83
	Points	.956	**.98**	.97	.98	.81	**.87**	**.84**
	Medals	.941	.96	.98	.97	.82	.69	.75
	Badges	.931	.93	**.99**	.96	**.92**	.51	.66

5 Conclusion

The paper shows a grassroots approach to understanding the gamification needs of students, and analysing how gamification elements impact on student engagement. Specifically, we analyse how gamification can be linked to student engagement in

CamaleOn, a Brazilian MOOC for highschool students being trained for higher education. This is a first step towards establishing *how to design better gamified environments* to support online education for underrepresented and underserved communities. This approach is best suited for MOOCs, which have a large amount of data, but do not necessarily obey any particular learning theory for gamification, and which need further improved to better serve their communities, *whilst 'running'*. Thus, this approach means measuring student impact in *real-time*, to be able to intervene at finer-granularity, e.g., in the design of the next run of a course. Further research we are already undertaking is analysing motivators for the student participation, as well as how student behavior can be attributed to their knowledge. Future work could involve re-evaluating such research questions and hypotheses with data from future academic years, to see how consistent the years are and to decrease the threat of external validity.

This proposal involves cutting-edge technologies and techniques evolving constantly, such as (for the areas of computer science only) user analytics, information retrieval, 'big data' processing, user profiling, social web information elicitation and usage, recommendations, semantic web representation and processing, various other technologies and techniques related to the emerging web science. Concretely, in the long-term, we expect to report advances, for instance, in user analytics visualisation techniques, adaptation and personalisation techniques combining content-based personalisation with social interaction. Moreover, in such massive online environments, new types of behaviors are taking place, and new behavioral patterns emerge, and this is where the expertise of our behavioral experts is essential.

References

1. Soni, S. and S. Dubey, *Towards Systematic Literature Review of E-learning.* International Journal of Scientific Research in Computer Science, Engineering and Information, 2018. **3**(3): p. 1389-1396.
2. Belanger, Y., J. Thornton, and R.C. Barr, *Bioelectricity: A quantitative approach--Duke University's first MOOC.* EducationXPress, 2013. **2013**(2): p. 1-1.
3. Kizilcec, R.F., C. Piech, and E. Schneider. *Deconstructing disengagement: analyzing learner subpopulations in massive open online courses*. in *Proceedings of the third international conference on learning analytics and knowledge*. 2013. ACM.
4. Hamari, J., J. Koivisto, and H. Sarsa. *Does Gamification Work?-A Literature Review of Empirical Studies on Gamification*. in *HICSS*. 2014.
5. Gibson, D. and P. Jakl, *Theoretical considerations for game-based e-learning analytics*, in *Gamification in education and business*. 2015, Springer. p. 403-416.
6. Antonaci, A., R. Klemke, and M. Specht. *The Effects of Gamification in Online Learning Environments: A Systematic Literature Review*. in *Informatics*. 2019. Multidisciplinary Digital Publishing Institute.
7. Khalil, Hanan & Ebner, Martin. *MOOCs Completion Rates and Possible Methods to Improve Retention - A Literature Review*. World Conference on Educational Multimedia, Hypermedia and Telecommunications, 2014.

8. Challco, G.C., et al., *Personalization of gamification in collaborative learning contexts using ontologies.* IEEE Latin America Transactions, 2015. **13**(6): p. 1995-2002.
9. de Santana, S.J., et al. *A quantitative analysis of the most relevant gamification elements in an online learning environment*. in *Proceedings of the 25th international conference companion on world wide web*. 2016. International World Wide Web Conferences Steering Committee.
10. Tenório, T., et al., *A gamified peer assessment model for on-line learning environments in a competitive context.* Computers in Human Behavior, 2016. **64**: p. 247-263.
11. Deterding, S., et al. *From game design elements to gamefulness: defining gamification*. in *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments*. 2011. ACM.
12. Muntean, C.I. *Raising engagement in e-learning through gamification*. in *Proc. 6th International Conference on Virtual Learning ICVL*. 2011.
13. Toda, A., et al. *A taxonomy of game elements for gamification in educational contexts: Proposal and evaluation*. in *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*. 2019. IEEE.
14. Fredricks, J.A., P.C. Blumenfeld, and A.H. Paris, *School engagement: Potential of the concept, state of the evidence.* Review of educational research, 2004. **74**(1): p. 59-109.
15. Lee, J.-S., *The relationship between student engagement and academic performance: Is it a myth or reality?* The Journal of Educational Research, 2014. **107**(3): p. 177-185.
16. Hew, K.F., C. Qiao, and Y. Tang, *Understanding student engagement in large-scale open online courses: A machine learning facilitated analysis of student's reflections in 18 highly rated MOOCs.* International Review of Research in Open and Distributed Learning, 2018. **19**(3).
17. Dutt, A., M.A. Ismail, and T. Herawan, *A systematic review on educational data mining.* IEEE Access, 2017. **5**: p. 15991-16005.
18. Lei, S., et al. *Towards understanding learning behavior patterns in social adaptive personalized e-learning systems*. 2013. AMCIS.
19. Hew, K.F., *Promoting engagement in online courses: What strategies can we learn from three highly rated MOOCS.* British Journal of Educational Technology, 2016. **47**(2): p. 320-341.
20. Papamitsiou, Z. and A.A. Economides, *Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence.* Journal of Educational Technology & Society, 2014. **17**(4): p. 49-64.
21. Romero, C. and S. Ventura, *Data mining in education.* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2013. **3**(1): p. 12-27.
22. Sani, S.M., A.B. Bichi, and S. Ayuba, *Artificial Intelligence Approaches in Student Modeling: Half Decade Review (2010-2015).* IJCSN-International Journal of Computer Science and Network, 2016. **5**(5).