Learning Analytics in MOOCs: Can Data Improve Students Retention and Learning?

Mohammad Khalil
Educational Technology
Graz University of Technology
Graz, Austria
mohammad.khalil@tugraz.at

Martin Ebner
Educational Technology
Graz University of Technology
Graz, Austria
martin.ebner@tugraz.at

Abstract: In order to study learners' behaviors and activities in online learning environments such as MOOCs, the demanding for a framework of practices and procedures to collect, analyze and optimize their data emerged in the educational learning horizon. Learning Analytics is the field that arose to comply with such needs and was denominated as a "technological fix to the long-standing problems" of online learning platforms (Knox, 2014). This paper discusses the significance of applying Learning Analytics in MOOCs to overcome some of its issues. We will mainly focus on improving students' retention and learning using an algorithm prototype based on divergent MOOC indicators, and propose a scheme to reflect the results on MOOC students.

Introduction

Massive Open Online Courses (MOOCs) provide massive amounts of data about learners and how they interact with an online learning environment. Since Siemens and Downes launched their first open online course back in 2008, MOOCs have been steadily spreading across the Internet (McAuley et al., 2010). Due to its openness, MOOC students vary in their heterogeneity such as age, gender, educational background and location. With an internet access, a student from anywhere in the world can access high quality courses such as the ones provided by Harvard University or Massachusetts Institute of Technology through the well-known MOOC provider, edX[1]. In addition to that, learners are not only limited to a single type path learning specialization. For instance, a student of social science major can attend a computer programming course, and this does not stop here. (S)he can get a certificate after successfully achieving all the required exams of that course.

According to the high number of MOOCs enrollees, a rich content of information can be stored in the databases of MOOCs servers. The accumulation of such data leads to what is so-called "Big Data". This term has become familiar in the recent years and it refers to "datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze." (Manyika et al., 2011). However, when noisy, unstructured and steep heterogeneous data are filtered, the examination and analysis becomes conceivable. In 2011, a special field called Learning Analytics has emerged after the growing needs to understand behaviour and attitudes of learners in online learning platforms and the needed advice in learning (Siemens, 2010). It is presumed that Learning Analytics is firmly related to other fields such as web analytics, educational data mining, academic analytics and business intelligence (Elias, 2011). The rapid growth of analytics on educational data is related to the advancements of computer technology and tools as well as the ease of acquiring educational data that are produced by online learning environments (Schön et al, 2011).

As a related topic, Knox (2014) as well as Khalil and Ebner (2015a) discussed the high potential of Learning Analytics when it is applied to educational datasets of MOOCs. Logging mouse clicks, tracking forums activity, time spent on tasks, quiz performance as well as login frequency of learners generate

1 http://www.edx.org (Last access, 04.12.2015)

consequential data which acts as a rich source of valuable knowledge for Learning Analytics researchers in the MOOCs environment. The promising goals of combining both fields can be linked with frameworks such as the one provided by (Khalil & Ebner 2015b) in which Learning Analytics can: a) predict grade, knowledge or performance of learners in MOOCs; b) intervene to control drop-out rate and detect students at-risk; c) recommend and personalize platforms to suit the needs of learners, teachers and institutions; d) reflect the outcome knowledge to improve future experience such as students experience; e) benchmark the weak points of learning environment systems, which is MOOC platforms in our case. Accordingly, this research study will aim attention at the implementation of Learning Analytics in the leading Austrian MOOC platform, iMooX (http://www.imoox.at).

This research work is a continuation of previous studies on this platform (Khalil & Ebner, 2015a; Khalil, Kastl & Ebner, 2016), and an edge to get into a deeper interval of analysis of learners data to examine the potential of Learning Analytics in order to: a) improve learning in MOOCs in general; b) enhance the completion rate in MOOCs; and c) study learner patterns and predict at-risk students. The research study will present the used methodology to establish the first algorithm prototype and introduce its results of testing students' information. Moreover, we will propose a Learning Analytics – MOOCs scheme to employ the feedback principle which will lead to accomplish our goals.

This publication is organized as follows: the next section lists the related work. Section 3 covers the research methodology. Section 4 gives an overview about the MOOC platform and the course structure. Section 5 shows our data analysis and the algorithm extraction. Finally, section 6 proposes the Learning Analytics – MOOC scheme.

Related Work

Predicting behaviour of students and recognizing who are at-risk is not a new topic. For instance, in 1985, Noel and Levitz described retention, linked it with students' success and stated that the course topic is a major factor to increase a successful completion rate (Noel & Levitz, 1985). It is noticed that subjects about attrition rate and completion rate are highly debated in the old time education and educational technology.

In relation to MOOCs, psychological and motivational factors have been proposed as phenomena to clarify the dropout rate and at-risk students (Khalil & Ebner, 2014). Santos et al. (2014) defined a threshold to detect dropout rate criteria. They found that forum activity is a reliable indicator to predict students who might drop in MOOCs. Further, Lackner, Ebner and Khalil (2015), adduced to revise MOOCs duration. They found that splitting them into two periods will increase the retention rate. Additionally, some studies pointed out to the number of assignments attempts as a factor to understand the completion rate (Coffrin et al., 2014). Similar to this research study, Balakrishnan and Coetzee (2013) compared behaviour of students who dropped out and who completed MOOCs by using the mathematical Markov chain. They found out that the students who do not login and check their course progress, usually dropout dramatically. At the end, the proposed Learning Analytics – MOOCs scheme was influenced by Course Signals (Arnold & Pistilli, 2012). The Application is made to provide a direct feedback to students through an intervention which is based on data analysis. Moreover, it works as a traffic light in which a student gets a red light color if (s)he is at-risk to fail and gets a green light color when (s)he is close to succeed in the course, while a yellow color indicates a potential problem of succeeding.

Research Methodology

This research study employed a procedure of collecting the data from two MOOCs presented in 2014 on the iMooX platform. The first course called Gratis Online Lernen, and abbreviated in this paper as GOL2014 (Ebner et al, 2015). The second course called Lernen im Netz, and is abbreviated as LIN2014 (Lackner & Kopp, 2014) The Learning Analytics application parses log files, which contains records of the students' activities on the learning platform, and filter them. After that, the filtered data are exported into a readable file and the results are clustered into two main categories: completed students, and those who successfully completed the course and did all the tasks. The second category is the students who dropped out during the course. The activities of each group are then assorted to categories and the prevailing behaviours are analyzed. At the end, we observed the differences in order to introduce an algorithm prototype based on weights to predict students who will drop during the course and proposed a scheme with a view to implement our results.

MOOC-Platform & Courses Overview

The Platform

iMooX is an online learning platform that follows the xMOOC family and was first introduced in 2013. It is the first MOOC platform in Austria and was founded by the collaboration of Graz University of Technology and University of Graz. The platform is enriched by Open Educational Resource courses and adheres to open education and lifelong learning paths (Neuböck, Kopp, & Ebner, 2015). iMooX supports the online pedagogy of offering courses to students on a weekly basis. Concurring with MOOCs forms of learning, it offers videos, multiple attempts quizzes, online discussions through forums and interactive learning objects. In addition, certificates are offered to the students who successfully pass all the required tasks at no cost.

Courses Description

Two courses were analyzed in this research study: Gratis Online Lernen and Lernen im Netz.

Gratis Online Lernen (GOL2014)

GOL2014 was an eight-week course that started in October 2014, offered by Graz University of Technology and provided in German language. The course focused on educating people through the internet and instructed them on how to do it. The workload was set to be 2 hours/week. There were 1012 participants in the course. 217 students successfully completed the course; therefore 21.5% was the completion rate in this MOOC.

Lernen im Netz (LIN2014)

LIN2014 was also an eight-week course that started in October 2014 till the mid of December 2014. The MOOC was offered by the University of Graz and taught in German language. The workload was set to be 5 hours/week and the topics were about using social media in the Open Educational Resources. 519 was the number of students who registered for the course, and the completion rate was 25%.

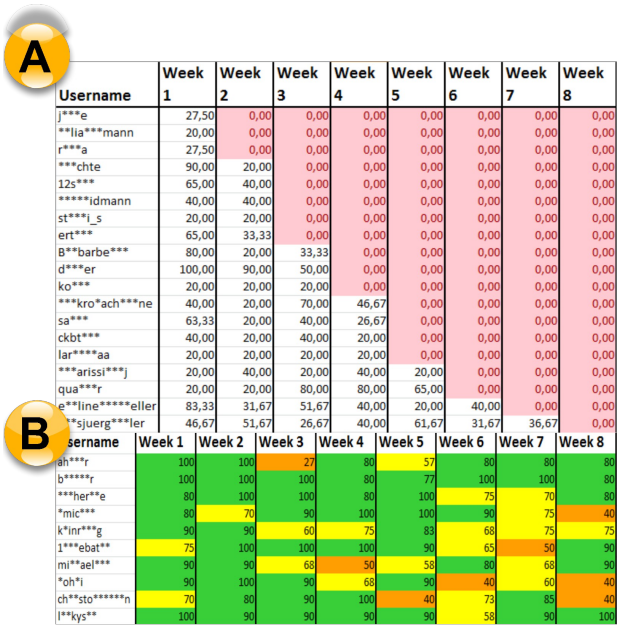
Analysis and Algorithm Extraction

In order to generate the prediction algorithm, an examination of GOL2014 and LIN2014 were scrutinized. Both of these two courses attracted the largest sample of students in iMooX platform in 2014. Basically, using the Learning Analytics approach, MOOC interactions have been filtered into indicators. These indicators are: a) quiz attempts; b) discussion forum readings; c) discussion forum writings; and d) login frequency. Each student profile was then dedicated with these indicators separately. In addition, all the collected data were distributed based on a weekly scale. Respectively, we calculated the total interactions for the completed students, who successfully finish MOOCs, and for the students who dropped and calculated the average.

Figure 1a, shows the behaviour of students who completed the course and the ones who dropped out in GOL2014. The left figure displays the average of all interactions in the whole course period, while the right figure shows the average of interactions from week3 to week6. The reason of analyzing that period is due to the stability of the dropout rate during that duration. Usually, students register in the first weeks to discover the material and the course, therefore the dropout rate in that period is quite high (Lackner, Ebner & Khalil, 2015). Moreover, the research study found that students, who stayed till week 5, have the potential to complete courses more often than in the earlier weeks. In the same fashion, figure 1b demonstrates the behaviour of both student types in the LIN2014 course. By observation of the line graph in both of the figures, students behaved nearly identical.

The figures show that the difference between quiz attempts is obvious between both of student types. Under that circumstance, quiz attempts factor has been pulled to be the second critical criteria in the algorithm.

Figure 1: The average of MOOC interactions for completed and dropout students. Top (a) GOL2014 course; bottom (b) LIN2014 course



A

Username	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8
j***e	27,50	0,00	0,00	0,00	0,00	0,00	0,00	0,00
lia*mann	20,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
r***a	27,50	0,00	0,00	0,00	0,00	0,00	0,00	0,00
***chte	90,00	20,00	0,00	0,00	0,00	0,00	0,00	0,00
12s***	65,00	40,00	0,00	0,00	0,00	0,00	0,00	0,00
*****idmann	40,00	40,00	0,00	0,00	0,00	0,00	0,00	0,00
st***i_s	20,00	20,00	0,00	0,00	0,00	0,00	0,00	0,00
ert***	65,00	33,33	0,00	0,00	0,00	0,00	0,00	0,00
B**barbe***	80,00	20,00	33,33	0,00	0,00	0,00	0,00	0,00
d***er	100,00	90,00	50,00	0,00	0,00	0,00	0,00	0,00
ko***	20,00	20,00	20,00	0,00	0,00	0,00	0,00	0,00
kro*achne	40,00	20,00	70,00	46,67	0,00	0,00	0,00	0,00
sa***	63,33	20,00	40,00	26,67	0,00	0,00	0,00	0,00
ckbt***	40,00	20,00	40,00	20,00	0,00	0,00	0,00	0,00
lar****aa	20,00	20,00	20,00	20,00	0,00	0,00	0,00	0,00
arissij	20,00	40,00	20,00	40,00	20,00	0,00	0,00	0,00
qua***r	20,00	20,00	80,00	80,00	65,00	0,00	0,00	0,00
e**line*****eller	83,33	31,67	51,67	40,00	20,00	40,00	0,00	0,00
sjuerg*ler	46,67	51,67	26,67	40,00	61,67	31,67	36,67	0,00

B

sername	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8
ah***r	100	100	27	80	57	80	80	80
b*****r	100	100	100	80	77	100	100	80
***her**e	80	100	100	80	100	75	70	80
*mic***	80	70	90	100	100	90	75	40
k*inr***g	90	90	60	75	83	68	75	75
1***ebat**	75	100	100	100	90	65	50	90
mi**ael***	90	90	68	50	58	80	68	90
*oh*i	90	100	90	68	90	40	60	40
ch**sto******n	70	80	90	100	40	73	85	40
l**kys**	100	90	90	90	90	58	90	100

Figure 2: Results of applying the algorithm on random samples from other courses. (a) Dropout students; (b) Completed students

However, the greatest difference gap among MOOC indicators is the forums reading activity. This explains several studies such as the one by Ezen-Can et al (2015), which have drawn attention to the significant effect of MOOCs forum interaction to improve learning and attrition.

Furthermore, the third remarkable difference is the login frequency which is noted as a decisive player in determining at-risk student, and this was highlighted in several studies like the one by Balakrishnan and Coetzee (2013). In contrast, the writings were not as efficient as readings; hence, the allocated weight for writing is the lowest.

Generally after all, each MOOC indicator was weighted and calculated based on the difference between the activity performance of students who dropped and the ones who completed in both of the two cases MOOCs. Subsequently, the algorithm predicts the retention percentage of each student and notify him/her at a peak point when (s)he is at-risk. We defined the weights according to their adequate significance to (W1, W2, W3, and W4), in which W1>W2>W3>W4. The following equation articulates the algorithm expression:

$$Success\ Rate\ (SR) = W1.Readings + W2.Quiz\_Attempts + W3.Login\_Frequency + W4.Writings$$

The first round testing showed promising results. In figure 2a, we see that students are dropping during the weeks when the algorithm generates low numbers. For instance, some students dropped when the score was below 40. However, some other students kept on track even when their score was 20. In such scenario, the student will be notified that (s)he is at-risk, and a proper reaction is required. Further, figure 2b displays a sample of students who completed the course. The orange tabs predict an at-risk situation, whereas the green records mean the student has the potential to complete the MOOC.

Learning Analytics – MOOCs Scheme

In this part, we are looking to attain the findings and achieve the feedback goal, so that learners can get the maximum benefits and be notified when their performance is in the danger area. Figure 3 shows the proposed Learning Analytics – MOOCs scheme. It is divided into five stages. The first stage starts when the data is generated through the activities of the learners. The log system records their quiz performance and attempts, discussion forum activity, their post and read count as well as committing login frequency to the database.

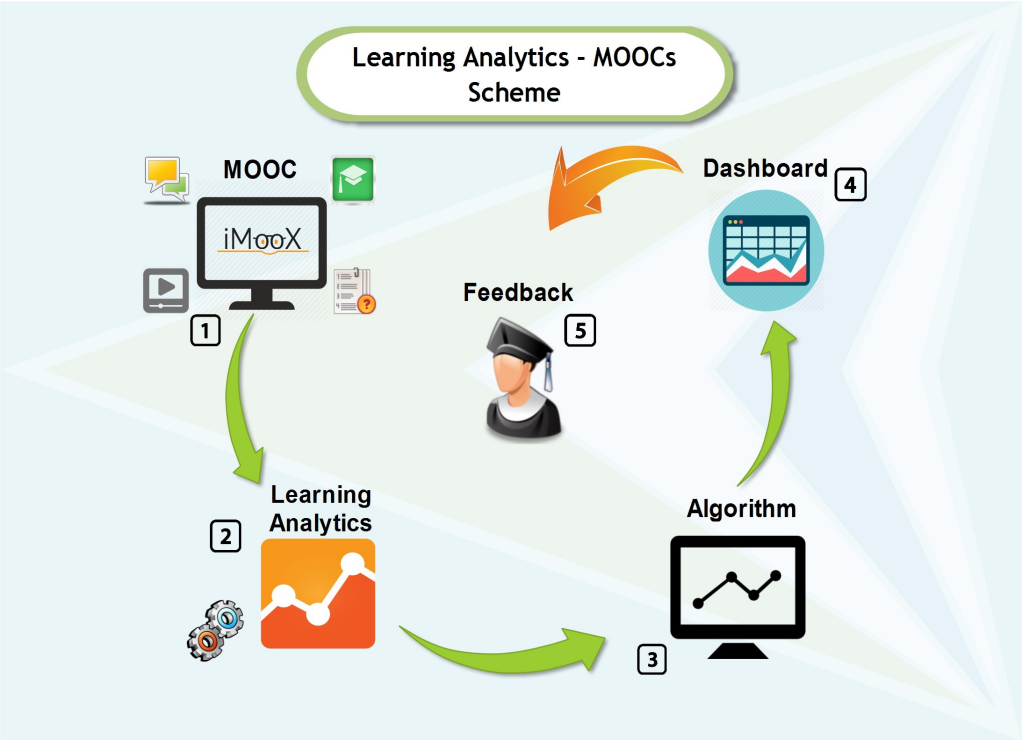


Figure 3: Learning Analytics – MOOCs Scheme

The second stage is where Learning Analytics server processes log files. It picks up keywords that help in determining students' interactions inside the bulk text of log files. In addition, it filters unstructured and duplicated data in order to be handled properly. In the third stage, the main activities of each student are parsed separately in order to support the algorithm procedures as discussed in the previous section. Upon calculating the weight of each interaction using mathematical operations, the results are formulated into a sequence of partitions for each student in furtherance of being dispatched to the next stage.

The fourth stage is where the collected, organized, operated interactions of learners' data are interpreted for the visualization part. The adopted method to show the results is the user dashboards. According to Verbert et al. (2014), dashboards support awareness, reflection, sense-making and ease learners to track their progress. The user interface should support feedback on activities and predict performance of students. The ideal of gamification could be presented in this proposed scheme using instruments such as: a progress bar or a colorful gauge. The aim is to boost students' motivation for learning and to sustain their interest in MOOCs. At this stage, the dashboard is intended to show a student's progress compared to other students. An indication of being behind the others, commensurate, or overhead will lead learners to react accordingly.

The last stage is the feedback section. As a consequence to the informative notification from the user dashboard, the awaited reaction is associated to a fruitful feedback. Timely, individual and empowering have been attributed as the needed qualities of feedbacks to get the awareness of students (Race, 2001). Thus, a student will be able to get updates about his/her performance on a weekly basis to ignite the learning competition and ambition. As a matter of fact, closing the feedback loop and enabling learners to react effectively is what pursued in our final scheme.

Conclusion

Learning Analytics is a field that promises to provide various solutions for online learning environments such as MOOCs. Despite that these courses are free to join, and attract a large volume of the community; the high ratio of dropout and the contradiction about its ability to create a true learning environment hinder the advancement of such a domain. Therefore, in this paper, we discussed the application of Learning Analytics on two MOOCs and proposed a brief literature review as well as described the walkthrough of tracking students' traces. Additionally, and in order to surpass MOOCs dilemmas, we implied our solution

for the purposes of predicting student at-risk and notify them beforehand by using an algorithm in order to increase retention rate, improve learning and study their behaviour. Moreover, we explained how we extracted the algorithm in details and proposed a Learning Analytics – MOOCs scheme that employ principles such as awareness and feedback.

In the meantime, the research team is working on testing more samples before the implementation stage and is developing compact visualizations that accommodate students and the institution needs. Additionally, we consider and respect users' privacy, therefore an implementation of a de-identification procedure (Khalil & Ebner, 2016), is undergoing to cover the Learning Analytics application and the Learning Analytics – MOOCs Scheme.

References

Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue: using learning analytics to increase student success. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (pp. 267-270). ACM.

Balakrishnan, G., & Coetzee, D. (2013). Predicting student retention in massive open online courses using hidden Markov models. *Electrical Engineering and Computer Sciences University of California at Berkeley*.

Coffrin, C., Corrin, L., de Barba, P., & Kennedy, G. (2014). Visualizing patterns of student engagement and performance in MOOCs. In*Proceedings of the fourth international conference on learning analytics and knowledge* (pp. 83-92). ACM.

Ebner, M., Schön, S. & Käfmüller, K. (2015). Inverse Blended Learning bei „Gratis Online Lernen" – über den Versuch, einen Online-Kurs für viele in die Lebenswelt von EinsteigerInnen zu integrieren. In: Digitale Medien und Interdisziplinarität. Nistor, N. & Schirlitz, S. (Hrsg). Waxmann, Medien in der Wissenschaft Bd 68. pp. 197-206

Elias, T. (2011). Learning analytics: Definitions, processes and potentials. Retrieved November 2015 from http://learninganalytics.net/LearningAnalyticsDefinitionsProcessesPotential.pdf.

Ezen-Can, A., Boyer, K. E., Kellogg, S., & Booth, S. (2015). Unsupervised modeling for understanding MOOC discussion forums: a learning analytics approach. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge* (pp. 146-150). ACM.

Khalil, H. & Ebner, M. (2014). MOOCs Completion Rates and Possible Methods to Improve Retention – A Literature Review. In *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2014* (pp. 1236-1244). Chesapeake, VA: AACE.

Khalil, M., & Ebner, M. (2015a). A STEM MOOC for school children—What does learning analytics tell us?. In *Interactive Collaborative Learning (ICL), 2015 International Conference on* (pp. 1217-1221). IEEE.

Khalil, M., & Ebner, M. (2015b). Learning analytics: principles and constraints. In *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications* (pp. 1326-1336).

Khalil, M., Kastl, C., & Ebner, M. (2016). Portraying MOOCs Learners: a Clustering Experience Using Learning Analytics. In *Proceedings of the European Stakeholder Summit on experiences and best practices in and around MOOCs (EMOOCS 2016),* Graz, Austria, pp.265-278.

Khalil, M. & Ebner, M. (2016). "De-Identification in Learning Analytics". Journal of Learning Analytics, 3 (1), pp. 129-138.

Knox, J. (2014). From MOOCs to Learning Analytics: Scratching the surface of the 'visual'. *eLearn*, *2014*(11), 3.

Lackner, E. & Kopp, M. (2014) Do MOOCs need a Special Instructional Design?. In: Proceedings of EDULEARN14 Conference 7th-9th July 2014, Barcelona, Spain, pp. 7138-7147

Lackner, E., Ebner, M., & Khalil, M. (2015). MOOCs as granular systems: design patterns to foster participant activity. *eLearning Papers*, *42*, 28-37.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity.

McAuley, A., Stewart, B., Siemens, G., & Cormier, D. (2010). The MOOC model for digital practice. Retrieved December

2015 from http://www.davecormier.com/edblog/wp-content/uploads/MOOC_Final.pdf.

Neuböck, K., Kopp, M., Ebner, M. (2015). What do we know about typical MOOC participants? First insights from the field, In*: Proceedings of eMOOCs 2015 conference, Lebrun, M., de Waard, I., Ebner, M., Gaebel, M., Mons, Belgium, pp. 183-190.*

Noel, L., & Levitz, R. (1985). Increasing student retention: New challenges and potential. In *L. Noel, R. Levitz, & D. Saluri (Eds.), Increasing student retention* (pp. 1-27). San Francisco, CA: Jossey-Bass.

Race, P. (2001). Using feedback to help students learn. *The Higher Education Academy, York*. Retrieved December 2015 from http://wap.rdg.ac.uk/web/FILES/EngageinFeedback/Race_using_feedback_to_help_students_learn.pdf.

Santos, J. L., Klerkx, J., Duval, E., Gago, D., & Rodríguez, L. (2014). Success, activity and drop-outs in MOOCs an exploratory study on the UNED COMA courses. In *Proceedings of the Fourth International Conference on Learning Analytics and Knowledge* (pp. 98-102). ACM.

Schön, M., Ebner, M. & Kothmeier, G. (2012). It's Just About Learning the Multiplication Table. In Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK '12), Simon Buckingham Shum, Dragan Gasevic, and Rebecca Ferguson (Eds.). ACM, New York, NY, USA, 73-81

Siemens, G. (2010). What are Learning Analytics? Retrieved December 2015 from http://www.elearnspace.org/blog/2010/08/25/what-are-learning-analytics/.

Verbert, K., Govaerts, S., Duval, E., Santos, J. L., Van Assche, F., Parra, G., & Klerkx, J. (2014). Learning dashboards: an overview and future research opportunities. *Personal and Ubiquitous Computing*, *18*(6), 1499-1514.