A MOOC on Educational Data Mining

Baker, R.S. Wang, Y., Paquette, L., Aleven, V., Popsecu, O., Sewall, J., Rose, C., Tomar, G., Ferschke, O., Zhang, J., Cennamo, M., Ogden, S., Condit, T., Diaz, J., Crossley, S., McNamara, D., Comer, D., Lynch, C., Brown, R., Barnes, T., Bergner, Y.

In this chapter, we describe a MOOC on educational data mining/learning analytics, *Big Data in Education* (referred to below as BDEMOOC in some cases). We will describe BDEMOOC's goals, its design and pedagogy, its content, and the research it afforded.

1. Big Data in Education – the course

    a. *Iteration 1: Coursera*

Big Data in Education was offered in its first version on the Coursera platform, as one of the inaugural courses offered by Columbia University. It was created in response to the increasing interest in the learning sciences and educational technology communities in learning to use educational data mining (EDM) methods with fine-grained log data. It was supported by initial investment from the Provost of Teachers College, Columbia University, and through the ongoing partnership between Teachers College and the Columbia Center for New Media Technology and Learning (CCNMTL).

The overall goal of this course was to enable students to apply core methods in educational data mining to answer education research questions and to drive intervention and improvement in educational software and systems. The course covered roughly the same material as a graduate-level course, Core Methods in Educational Data Mining, at Teachers College Columbia University. However, most topics were covered in less depth than in that course, and there was less scope for students to develop creative solutions to problems than in that earlier course. BDEMOOC's first iteration began on October 24, 2013. It officially ended on December 26, 2013, but the course remained open after that point.

The weekly course was comprised of lecture videos and 8 weekly assignments (offered through the quiz functionality of Coursera). Most of the videos contained in-video quizzes that did not count toward the final grade. All the weekly assignments were automatically graded and involved numeric input or multiple-choice questions. In each assignment, students were asked to conduct an analysis (or set of analyses) on a data set provided to them and answer questions about the data set. The original design goal was for assignments to be internally cumulative (i.e., each step built on the previous step), but this goal was not fully achieved. Coursera did not support assignments that presented the steps to the student, step-by-step, but required assignments to be shown in their entirety from the start, making it necessary to avoid giving away the answers to previous steps within later steps. In order to receive a grade, students had to complete each assignment within two weeks of its release. Students typically were given up to three attempts for each assignment, but in some cases additional attempts were offered when problems had bugs in them. The highest score the student achieved on the assignment was counted as their final grade for that assignment.

The course had a total enrollment of 45,268 during its official run as a course (an additional 20,316 joined and accessed the course after the official end date). A smaller number actively participated: 13,314 students watched at least one video; 1,242 students watched all the videos; 1,380 students completed at least one assignment; and 710 made a post in the weekly discussion sections. Of those with posts, 426 students completed at least one class assignment. 2,776 completed a pre-course survey; we will discuss data from this survey below.

A total of 638 students completed the online course and received a certificate. Many students successfully completed the course and earned certificates without ever posting in the discussion forums.

b. Iteration 2 -- edX

The second iteration of Big Data in Education was offered on the edX platform, with support from grant funding from the National Science Foundation, and through the ongoing partnership between Teachers College and the Columbia Center for New Media Technology and Learning (CCNMTL).

The lecture content of this second iteration of the course was largely the same as the first iteration. A few lectures were modified based on new developments in the field, and new information provided by colleagues or students in the course – for example, advice from a colleague (Radek Pelanek) led to changing the discussion of the virtues of the RMSE metric versus the MAE metric. A small set of corrections was also made based on errors discovered in the first iteration. For example, one set of images in the first iteration had been incorrectly attributed to the wrong author.

The larger changes between the first and second iterations of BDEMOOC involved new types of assignments, added with the goal of improving the scaffolding of complex problem-solving processes and conceptual learning. Three new types of activities were added:

- Cognitive tutor-based assignments, replacing the earlier Coursera quiz-based assignments. These were offered through the CTAT (Cognitive Tutor Authoring Tools) platform (Aleven et al., 2009)
- Collaborative chat activities, offered through the Bazaar platform (Ferschke et al., 2015a, 2015b)
- Tool walkthroughs, offered as pdf files

BDEMOOC's second iteration began on July 1st, 2015. It officially ended on August 26th, 2015, but the course remained open after that point.

The video lectures are also available on an ongoing basis from Professor Ryan Baker's webpage at http://www.columbia.edu/~rsb2162/bigdataeducation.html.

The course had a total enrollment of 10,348 during its official run as a course. A smaller number actively participated; we have not yet received data on usage, but anticipate receiving it in the next month, and will include more detail on usage numbers in the final version of this chapter.

We do currently know that 251 students completed at least one assignment, and that 113 students in total completed the online course and received a certificate. Many students successfully completed the course and earned certificates without ever posting in the discussion forums.

CTAT:

The second iteration of BDEMOOC had assignments developed in CTAT, the Cognitive Tutor Authoring Tools. CTAT supports the rapid authoring of intelligent tutoring system activities that offer step-by-step guidance for complex problem solving activities (Aleven et al., 2009). For the second iteration of BDEMOOC, CTAT was integrated with the edX platform to provide multi-step activities, with:
a) Hint messages, offered at almost every step, guide students through the thinking processes necessary to produce the correct answer
b) Buggy Messages provide immediate, detailed feedback when students provide a wrong answer that indicates a known misconception

Students can voluntarily choose whether to access the hint messages and how many hint messages they would like to view. The assignments were designed to guide the student through a step-by-step process; the next problem step or question does not appear until the student successfully completes the current question.

Further detail on the CTAT assignments, and how they were integrated into edX, is provided in (Aleven et al., 2015).

Bazaar:
In addition to the CTAT assignments, the second iteration of BDEMOOC also had collaborative chat activities supported by the Bazaar tool (Ferschke et al., 2015a, 2015b). Like CTAT, Bazaar was integrated into the edX platform. In this collaborative activity, students enter the chat room where they are paired with one or more partner students to discuss topics related to the learning materials offered in the corresponding week. A conversational computer agent ("Virtual Ryan") facilitated the conversation.

Compared to the first iteration of this course, incorporating interactive activities enabled the course to provide students with better-scaffolded learning experiences in the following two areas:
a. Enrichment of learning activities – The Bazaar activities provide students with novel opportunities to discuss what they have learned and interact with their peers beyond regular lecture videos and discussion forums.
b. Support for in-depth consideration of class issues – Virtual Ryan scaffolded students in discussing the concepts brought up in class lectures in greater depth than is typically seen in MOOC discussion forums. In these discussions, explicit connections are made to student interests, through asking students to identify a real-world challenge or goal they are interested in, and then guiding the students to discuss concepts in terms of those challenges.

Walkthrough:
Another addition in the second iteration of the course was a Walkthrough file to introduce the RapidMiner tool required to complete the majority of the graded assignments. RapidMiner is an open-source software platform for machine learning and data mining (RapidMiner Studio, 2015).

Since many learners had little or no experiences using RapidMiner prior to taking this course, a walkthrough file containing basic beginning steps such as how to import a dataset in RapidMiner and create a cross-validated prediction model was included as part of the first week's learning materials.

iii. Course content

In both its iterations, BDEMOOC covered the following topics listed by week below. The activities listed pertain to the second iteration of the course.
(In all cases, further detail is given in the actual BDEMOOC content itself, available at http://www.columbia.edu/~rsb2162/bigdataeducation.html)

Week 1: Prediction Modeling

In prediction modeling, the goal is to develop a model which can infer a single aspect of the data (the predicted variable, similar to dependent variables in traditional statistical analysis) from some combination of other aspects of the data (predictor variables, similar to independent variables in traditional statistical analysis). The course covered algorithms for classification (predicting a binary or categorical/polynomial variable) and regression (predicting a number) that have been found to be useful within educational data sets. The strengths and weaknesses of a small but representative set of algorithms was discussed. Students completed a walkthrough showing them how to import and model data in RapidMiner, and then completed a CTAT assignment that involved comparing different algorithms and validation approaches on the same data set. They then used Bazaar to discuss the CTAT assignment, and how prediction modeling methods might be useful in their own work.

Week 2: Diagnostic Metrics

In the second week of the course, we discussed diagnostic metrics that were relevant for classification and regression models, covering accuracy, kappa, A'/AUC ROC (the area under the Receiver Operating Characteristic Curve)/Wilcoxon, precision, recall, correlation and RMSE (Root Mean Squared Error). (For definitions of these terms, see the course itself). We also discussed the role played by detector confidence – the detector's estimate of the probability that a specific prediction is correct – and why it is important to preserve the information available in detector confidence. Issues of how cross-validation can be used to assess and avoid over-fitting were discussed. The week's lectures concluded with a discussion of the different ways prediction models can be valid or invalid. Students completed a CTAT assignment that involved computing a set of different diagnostic metrics on two data sets, and then participated in a Bazaar activity where they discussed whether a model for a specific goal is valid and generalizable.

Week 3: Behavior Detection and Feature Engineering

The third week of the course focused on a specific type of prediction modeling, behavior detection – where specific student behaviors of interest are identified, ground truth data is collected, and then a model is built to automatically recognize those behaviors of interest. Particular focus was given to the issue of feature engineering, the distillation of raw or basic log

data into more complex and meaningful features that can be used as the basis for an automated detector of student behavior. Work on automated feature generation and selection was also discussed. The week's lectures concluded with a discussion of the relationship between knowledge engineering and data mining, with consideration of the continuum from building a model fully by hand, to distilling features by hand and using data mining to select and combine them, to building a model in a fully automated fashion. Students completed a CTAT assignment that covered the aggregation of data across multiple grain-sizes for use in behavior detection, and participated in a Bazaar activity that covered the early steps of feature engineering, including brainstorming potential features, selecting which ones to build, and writing out a feature definition/specification.

Week 4: Knowledge Inference

The fourth week of the course focused on a special type of prediction model: latent knowledge estimation models that conduct knowledge inference. In latent knowledge estimation, a student's knowledge of specific skills and concepts is assessed by their patterns of correctness on those skills (and occasionally other information as well). The models used in online learning typically differ from the psychometric models used in paper tests or in computer-adaptive testing, because with an interactive learning application, the student's knowledge is continually changing. A wide range of algorithms exist for latent knowledge estimation; the fourth week of the course focused on the two most popular: Bayesian Knowledge Tracing (BKT -- Corbett & Anderson, 1995) and Performance Factors Analysis (PFA -- Pavlik, Cen, & Koedinger, 2009). For comparison, there was also discussion of the popular psychometric approach, Item Response Theory (Embretson & Reise, 2013). In the CTAT assignment, learners built a Bayesian Knowledge Tracing model, and in the Bazaar assignment, learners discussed whether BKT and PFA were appropriate for a range of potential application contexts and domains.

Week 5: Relationship Mining

The fifth week of the course changed track, and focused on the varied methods of relationship mining, defined as analyses with the goal of discovering relationships between variables in a data set with large numbers of variables (Baker & Siemens, 2013). Lectures covered correlation mining, causal mining, association rule mining, sequential pattern mining, and network analysis. In correlation mining, the goal is to find positive or negative linear correlations between variables (using post-hoc corrections or dimensionality reduction methods when appropriate to avoid finding spurious relationships). In causal data mining, the goal is to find whether one event (or observed construct) was the cause of another event (or observed construct). In association rule mining, the goal is to find if-then rules of the form that if some set of variable values is found, another variable will generally have a specific value. In sequential pattern mining, the goal is to find temporal associations between events. In network analysis (sometimes called social network analysis after its primary application), models are developed of the relationships and interactions between individual actors, as well as the patterns that emerge from those relationships and interactions. In the CTAT assignment, learners conducted correlation mining, and considered when statistical validation of correlations is appropriate and inappropriate, and when post-hoc controls are appropriate. In the Bazaar activity, learners considered whether and

how each type of relationship mining could be applied to specific data sets related to their personal interests.

Week 6: Visualization

In the sixth week of the class, the lectures discussed a range of visualizations relevant to educational data, including both visualizations of broad applicability (scatter plots versus heat maps), and visualizations more characteristic of educational data (learning curves, moment-by-moment learning graphs, parameter space maps, and state space networks) – although each of these visualizations also has applications in other domains. In each case, lectures gave examples from published EDM research. The CTAT assignment covered a topic from the previous week, sequential pattern mining, leading students through the process of creating sequential patterns and studying how the setup of these algorithms impacts their results. The Bazaar activity focused on visualization, and led students through discussing how published visualizations succeeded or failed in communicating core information.

Week 7: Structure Discovery

The seventh week of the class focused on algorithms for structure discovery, approaches that attempt to find structure in the data without an a priori idea of what should be found. Lectures discussed clustering (methods for finding structure between data points), factor analysis (methods for finding structure between variables), and methods for finding structure in student knowledge. In terms of student knowledge, the lectures discussed both q-matrix methods (which find mappings between items and skills), and methods that infer hierarchy in student knowledge. The CTAT assignment led students through the process of deciding how many clusters to search for, and considering how different variables influence the results of a clustering approach. The Bazaar activity asked students to consider when clustering is an appropriate approach to use, and which clustering algorithm would be appropriate for a specific problem.

Week 8: Advanced Topics

The eighth and final week of the course covered a variety of methods that did not fit cleanly in any of the other seven weeks, including discovery with models (where the results of one data mining analysis are utilized within another data mining analysis), text mining, and Hidden Markov models (an approach for attempting to study the change in an agent's state over time). The CTAT assignment for week 8 gave students data from the first iteration of BDEMOOC and asked them to construct the course's social network and make inferences about the course's participants from their interactions with each other. The Bazaar activity asked course participants to discuss their goals for using EDM in their careers going forward, what approaches might be useful to their goals, and to reflect on what they had learned in the course.

2. Research on BDEMOOC

Beyond BDEMOOC's possible benefits to the tens of thousands of students who have participated in it in one fashion or another, BDEMOOC has supported researchers in the

educational data mining research space. In this section, we discuss some of the research that BDEMOOC has facilitated.

a. Post-MOOC participation in the EDM community of practice

One key question is how participation in a professionally-oriented advanced MOOC such as BDEMOOC influences longer-term student participation in careers involving EDM. A survey was given to students at the end of the first iteration of BDEMOOC. At the end of the course, more than 80% of respondents (among a limited set of 536 respondents, not all of whom completed the course) plan to use the skills taught in this MOOC in their career (Wang, Baker, & Paquette, 2014).

Going forward from self-reported intent to use the skills to actual community participation, we have found that 35 students who registered for the MOOC joined the International Educational Data Mining Society in the first several months after the course started. Preliminary analyses indicated that these 35 students disproportionately completed the course; 20% of students who joined the society completed the course. By comparison, only 1.3% of the students who did not join the society completed the course (χ^2 (1) = 97.438, $p < 0.001$) (Wang, Paquette, & Baker, 2015). As such, course completion is a strong sign of the type of interest that leads students to join the associated scientific community.

Although 35 is a small number in comparison to the total enrollment, it only reflects one small aspect of community participation. Another indicator of community participation comes from Danielle McNamara's keynote address at the International Conference on Learning Analytics and Knowledge in 2015. During her keynote, she asked members of the audience who had enrolled in BDEMOOC's first iteration, and around 40% of the people in the room raised their hands. Further data on learner participation in the EDM and LAK scientific communities is also in the process of being collected, in order to better understand what role (if any) this MOOC played in the role of the development of the field. If this project is successful, it may contribute to our understanding of what it means for a MOOC to succeed, and how the effectiveness of a MOOC can be assessed more comprehensively.

ii. Predicting early stop-out

In recent unpublished work, Zhang and colleagues (in preparation) have taken data from BDEMOOC and attempted to infer how long students will persist in BDEMOOC from their performance in the first two weeks of the course. They analyzed data from early-week assignments, forum participation (including passive reading), and watching and downloading videos. They found that it is possible to derive a model from these features that predicts approximately 25% of the variance in how long students persist in the course (after the second week). This model may be usable in later variants of the course for helping to support students who are relatively more likely to stop participating in the course (cf. Whitehill et al., 2015).

iii. Predicting drop-out from forum participation

Crossley and colleagues (2015) investigated the relationship between student language on the discussion forum and course completion. This research, complementary to other analyses of

MOOC forum data which looked at the presence of specific words (Wen et al., 2014), used natural language processing (NLP) to examine whether the types of language in the discussion forum of an educational data mining MOOC is predictive of successful class completion.

The analysis was applied to a subsample of 320 students who completed at least one graded assignment and produced at least 50 words in discussion forums. The findings indicate that the language produced by students can predict substantially better than chance (cross-validated Cohen's Kappa = 0.379) whether students complete the MOOC. While many students did not use the discussion forums, students who participated more frequently in the forums, who used a greater range of vocabulary in the discussion forum, and who used more concrete and sophisticated words, were more likely to complete the MOOC. These results suggest that NLP can help us to better understand student retention in MOOCs.

iv. Participation in sub-communities

Social network analysis has also been adopted in analyzing forum data from the course. Brown and colleagues (2015) examined social network graphs drawn from forum interactions in BDEMOOC to identify natural student communities and characterize them based on student performance and stated preferences. They found that students' grades were significantly correlated with their most closely associated peers in the network. Their findings suggest the students are forming communities that are homogeneous with respect to course outcomes.

v. Negativity towards instructors

Student negativity can take several forms in MOOCs, but includes (most notably) hostile or insulting comments towards the instructor or other students (several colorful examples of this can be found in Comer et al., 2015). Negativity towards MOOC instructors has been little-studied, but is thought to play an important role in instructor disengagement (Comer et al., 2015), where instructors reduce or cease their active participation in their own course, or decide not to teach another iteration of their MOOC. Negativity has been part of MOOCs since the beginning, influencing the design of the second iteration of the first MOOC (personal communication, George Siemens). BDEMOOC provided a venue for further study of negativity towards instructors (Comer et al., 2015).

Among consistent forum contributors in BDEMOOC, nine participants displayed repeated negativity toward the instructor. Although these numbers represent a tiny percentage of over 48,000 registered students, they accounted for a disproportionate number of negative comments. A small number of outspoken students can create a substantial negative experience for instructors. Of these nine consistently negative individuals, four also responded to a pre-course survey on their motivations (cf. Wang & Baker, 2015). This rate of response (44.44%) was much higher than the rest of the class's response rate (2.9%), $\chi 2$ (df=1)= 55.31, $p<0.0001$ (Wang, Paquette, & Baker, 2014). However, no motivational survey items differentiated these negative individuals from other students in the class. Interestingly, all of the consistently negative students appeared to be male (according to either the pre-course survey or their choice of name on the forums).

Further qualitative analysis on negativity in BDEMOOC demonstrates the multifaceted nature of negativity in MOOCs and the importance of finding ways to mitigate negativity and support instructors who experience it in their courses. Much of the negativity encountered during the course was related to elements of course design inherent in the platform or related to design choices made prior to the beginning of the course. Most of the labor and instructional time invested by an instructor occurs prior to a MOOC's launch, leaving the instructor with limited capability to make changes that can help address this source of negativity during the course itself. This suggests that managing negativity should be integrated into the process of course design and development—perhaps with an eye towards creating design principles for next-generation MOOCs that reduce negativity and mitigate its effects (see Comer et al., 2015 for discussion of some design possibilities). It is important to note that negativity can inform instructors and staff about problems with the course that can and should be addressed. As such, it is not necessarily optimal (or possible) to eliminate negativity entirely, but it is important to make sure that negativity does not result in instructor disengagement.

One factor which may have increased the degree of negativity and limited the potential for response by the instructor or other course (or platform) staff was the open nature of the MOOC, where students did not have to pay money and were not attempting to obtain course credit leading toward earning a degree. Some degree of the negativity seen here may be particular to MOOCs; in a regular course, a disruptive or abusive student could ultimately be removed from the course or referred to university disciplinary authorities. In addition, the instructor's ability to assign grades in a traditional course likely restrains student negativity to some degree. Even if an instructor removed a student from a course, in an open MOOC there would be little to prevent the student from creating a new identity, rejoining the course, and resuming the negative behavior. As such, instructors in MOOCs have considerably fewer options for dealing with negativity than instructors in for-credit online courses.

Interestingly, however, very little negativity was seen in the second iteration of BDEMOOC compared to the first iteration. It is not clear whether this was due to the greater degree of polish in BDEMOOC's second iteration; it is probably not due to edX students being more positive than Coursera students, as Baker participated in another edX MOOC, *DALMOOC*, and observed considerable negativity towards instructors there.

v. Understanding the costs of MOOCs

During the preparation of the first iteration of BDEMOOC, Baker collected extensive data on all of his work in the MOOC. Most of this was reported at a five-minute grain-size. These data were used in turn by Hollands and Tirthali at the Center for Benefit-Cost Studies of Education to study the costs of MOOC development and delivery, and how the instructor's time is spent across various MOOC-related activities (Hollands & Tirthali, 2014). Hollands and Tirthali were able to determine that BDEMOOC took 176 hours of instructor time to develop and deliver, with planning and bureaucracy taking almost as much time as recording video lectures. They estimated that the costs of offering this MOOC were around $39,000 but found that costs are highly sensitive to the salary level of the instructor and would be higher if an external video production team were hired. Hollands and Tirthali also reported that Baker's scientific

productivity (in terms of journal articles, book chapters, and conference paper proposals submitted) was reduced during the time he was developing and delivering the MOOC.

vii. Understanding MOOC learner motivation

In order to better understand the learners who take MOOCs, a survey of MOOC learners' motivations was conducted. This survey was then correlated with course completion (Wang & Baker, 2015). The results showed that course completers tended to be more interested in the course content, whereas non-completers tended to be more interested in MOOCs as a platform. Contrary to initial hypotheses, however, completers were not found to differ from non-completers in terms of mastery-goal orientation or general academic efficacy. However, students who completed the course tended to be more confident that they would complete the course, from the beginning.

3. Conclusion

In this chapter, we have described a MOOC on educational data mining/learning analytics, *Big Data in Education.* We have described BDEMOOC's goals, its design and pedagogy, its content, and the research it afforded. The first iteration of BDEMOOC was taught in a fairly standard fashion, on the Coursera platform – in its second iteration, BDEMOOC was ported to the edX platform and extended to include additional activities such as collaborative chat and step-by-step problem-solving.

BDEMOOC has supported a number of research projects, making it one of the more thoroughly-studied MOOCs. Researchers have used BDEMOOC to study post-MOOC participation in scientific communities of practice, drop out and early stop-out, MOOC learners' motivations, negativity towards instructors, and the costs of producing a MOOC. These research projects are mostly still in their early stages, as is most of the research around MOOCs, but point to the potential of collaborations between researchers with questions on MOOCs and a highly engaged course development team.

In future years, BDEMOOC is expected to continue to run as long as there is substantial student demand for its content and learning experience. Future efforts will include attempts to improve the coordination of student participation in collaborative chat, and further iterative refinement to the content and other learning activities. We anticipate continued research and development efforts to keep BDEMOOC up to date and to keep it a useful service to the broader community of scientists and practitioners in this area.

Acknowledgments

We would like to thank Fiona Hollands for useful comments and suggestions. We would also like to thank NSF #DRL-1418378.

References:

Aleven V., McLaren B.M., Sewall J., Koedinger K.R. (2009). A new paradigm for intelligent tutoring systems: Example-Tracing tutors. International Journal of Artificial Intelligence in Education 19, 105-154.

Aleven, V., Sewall, J., Popescu, O., Xhakaj, F., Chand, D., Baker, R., ... & Gasevic, D. (2015). The Beginning of a Beautiful Friendship? Intelligent Tutoring Systems and MOOCs.

Brown, R., Lynch, C., Wang, Y., Eagle, M., Albert, J., Barnes, T., Baker, R., Bergner, Y., McNamara, D. (2015) Communities of Performance & Communities of Preference. *Proceedings of the Graph Analytics Workshop at the International Educational Data Mining (EDM) Conference*.

Crossley, S., McNamara, D., Baker, R.S., Wang, Y., Paquette, L., Barnes, T., Bergner, Y. (2015) Language to Completion: Success in an Educational Data Mining Massive Open Online Course. *Proceedings of the 8th International Conference on Educational Data Mining*, 388-391.

Corbett, A.T., Anderson, J.R., 1995. Knowledge tracing: Modeling the acquisition of procedural knowledge. User Modeling and User-Adapted Interaction, 4 (4), 253-278.

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. Mahwah, NJ.: Lawrence Erlbaum Associates

Hollands, F. M. & Tirthali, D. (May, 2014). *MOOCs: Expectations and reality.* Center for Benefit-Cost Studies of Education. New York, NY: Teachers College Columbia University. Retrieved from http://cbcse.org/wordpress/wpcontent/uploads/2014/05/MOOCs_Expectations_and_Reality.pdf

Ferschke, O., Howley, I., Tomar, G., Yang, D., Rosé, C. P. (2015a). Fostering Discussion across Communication Media in Massive Open Online Courses *Proceedings of Computer Supported Collaborative Learning*

Ferschke, O., Yang, D., Tomar, G., & Rosé, C. P. (2015b). Positive Impact of Collaborative Chat Participation in an edX MOOC. In *Artificial Intelligence in Education* (pp. 115-124). Springer International Publishing.

Kizilcec, R. F., Piech, C., & 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* (pp. 170-179).

Pavlik, P., Cen, H. and Koedinger, K.R. (2009). Learning Factors Transfer Analysis: Using Learning Curve Analysis to Automatically Generate Domain Models. In *Proceedings of the 2nd International Conference on Educational Data Mining*, 121-130.

RapidMiner Studio. (2015). Retrieved from https://rapidminer.com/products/studio/

Wang, Y. Baker, R. (2015) Content or Platform: Why do students complete MOOCs? *MERLOT Journal of Online Learning and Teaching, 11* (1), 17-30.

Wang, Y.E., Paquette, L., Baker, R. (2015) A Longitudinal Study on Learner Career Advancement in MOOCs. *Journal of Learning Analytics, 1* (3), 203-206.

Wen, M., Yang, D., & Rose, C. P. (2014). Linguistic reflections of student engagement in massive open online courses. In *Proceedings of the International Conference on Weblogs and Social Media*.

Whitehill, J., Williams, J., Lopez, G., Coleman, C., Reich, J. (2015) Beyond Prediction: Towards Automated Intervention in MOOC Student Stop-out. *Proceedings of the 8th International Conference on Educational Data Mining*, 171-178.