M-CAFE 2.0: A Scalable Platform with Comparative Plots and Topic Tagging for Ongoing Course Feedback

Mo Zhou[1], Sanjay Krishnan[2], Jay Patel[2], Brandie Nonnecke[3]
Camille Crittenden[3], Ken Goldberg[1,2]

[1]UC Berkeley
IEOR Department
{mzhou, Goldberg}@berkeley.edu

[2]UC Berkeley
EECS Department
{sanjaykrishnan, patel.jay}@berkeley.edu

[3]UC Berkeley
CITRIS Connected Communities Initiative
{nonnecke, ccrittenden}@berkeley.edu

ABSTRACT
M-CAFE 2.0 is an online and mobile platform that uses collaborative filtering to collect and organize student feedback each week throughout a MOOC or an on-campus course to facilitate mid-course corrections by instructors. M-CAFE 2.0 encourages students to assess course content, structure, and suggestions provided by their peers. It requires minimal extra effort from instructors and is anonymous and separate from all student records. We present results from three pilot studies from on-campus undergraduate courses with 1,211 evaluations and 5,221 peer-ratings from 169 students. Results suggest that comparative plots of past ratings, topic tags and peer-to-peer anonymous suggestion evaluations are valuable in promoting credible and diverse course evaluation. M-CAFE 2.0 is available at m-cafe.org.

Author Keywords
Course Evaluation; Student Evaluation of Teaching; Collaborative Filtering; MOOCs.

CSS Keywords
• Human-centered computing~Collaborative filtering
• Human-centered computing~User interface design
• Applied computing~Computer-assisted instruction
• Applied computing~E-learning

SIGITE'17, October 4–7, 2017, Rochester, NY, USA

ACM ISBN 978-1-4503-5100-3/17/10◄$15.00
https://doi.org/10.1145/3125659.3125681

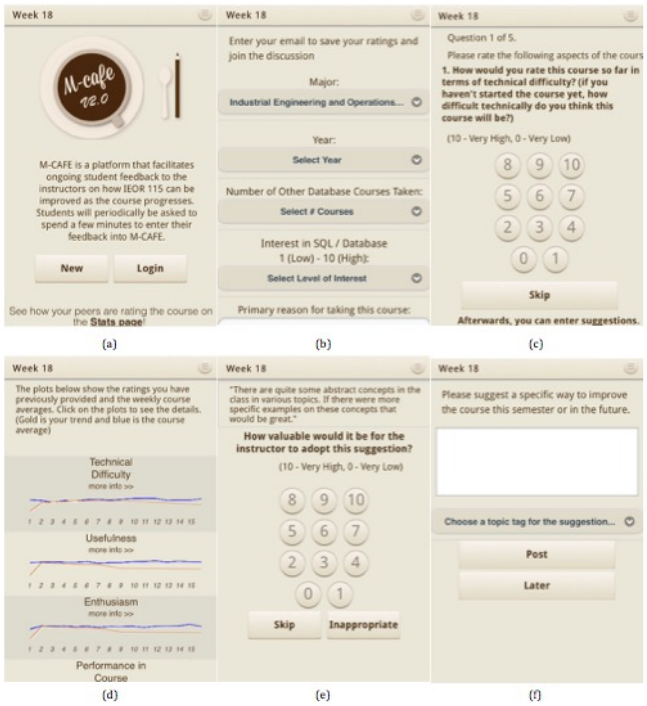

Figure 1: Screenshots of the M-CAFE 2.0 interface, compatible with desktop and mobile devices: a. the landing screen; b. new users register using their email address and provide demographics; c. users rate the course aspects (Course Difficulty, Course Usefulness, Self-Enthusiasm, Self-Performance and Homework Effectiveness); d. the comparative plots of past ratings; e. users rate the value of a peer-provided suggestion; f. users enter their own course improvement suggestions.

1. INTRODUCTION
In recent years, on-campus courses are experiencing growing class sizes, making individualized student-instructor interaction challenging. In the meantime, massive open online courses (MOOCs) have received widespread attention and excitement since 2012. But studies find that MOOCs have an extremely high dropout rate, due to limited interaction with instructors and peers, insufficient

academic background and personal stress [6, 14, 20, 22, 23]. Therefore, more specific and frequent feedback from learners may help instructors identify problems in course content and provide timely responses for both MOOCs and large on-campus courses.

In 2014, we developed the MOOC Collaborative Assessment and Feedback Engine 1.0 (M-CAFE 1.0) to collect ongoing student feedback [24].

In this paper, we introduce M-CAFE 2.0, an enhanced version of M-CAFE 1.0 with comparative plots of ratings on Course Difficulty, Course Effectiveness, Self-Enthusiasm, Self-Performance and Homework Effectiveness and topic tagging for discussion suggestions centered around the question "In what specific way would you improve this course this semester or in the future?" Displaying comparative plots serves as a baseline for future ratings and enables students to quickly track their progress. By comparing self-ratings to the course average, students see where they stand among their peers. To assess the performance of M-CAFE 2.0 on traditional on-campus courses, we adopted it in two large on-campus courses at UC Berkeley in 2016 taught by Professor Ken Goldberg.

The paper is structured as follows: we first evaluate the current state of course evaluation. Then we introduce M-CAFE 2.0 and describe its features and the improvements from M-CAFE 1.0. We report results from three pilot studies with 169 students.

2. RELATED WORK

2.1 Student Evaluation of Teaching

Student evaluation of teaching (SET) is widely used to evaluate instructor's teaching effectiveness. SET usually occurs at the end of the semester when students are asked to rate different aspects of the course on a numerical scale (usually from 1 to 10 "Very Low" to "Very High"). Many studies question the validity of naively aggregating quantitative ratings from SET [1, 3, 4, 11, 13, 17]. Stark and Freishtat [18] state that "SET are ordinal categorical variables and comparing an individual instructor's average to department average is meaningless, since there's no reason to believe that the difference between 3 and 4 means the same thing as the difference between 6 and 7 and that the difference between 3 and 4 means the same thing to different students." McCullough & Radson further introduce an alternative method to analyze SET data that uses categorical proportions instead of assigning a score to each category and numerically aggregating the results to robustly evaluate teacher's performance [12]. In contrast, Khong conducted a recent study of 200 students to suggest that the SET is a valid instrument in evaluating teaching effectiveness by measures such as internal consistency and correlations [7]. Surgenor suggests that SET can be valuable to measure dedication to teaching and improvement and to promote quality learning [19]. Furthermore, he identifies the need for easily obtainable feedback on modules. M-CAFE 2.0 is a tool to help instructors improve teaching effectiveness by encouraging and examining feedback on a weekly basis. To ensure comparable SET responses, M-CAFE 2.0 displays the average ratings for the past weeks as a benchmark for future ratings.

2.2 Collective Intelligence

Online discussion platforms such as Piazza and stackExchange are popular collective intelligence sites. Most MOOCs and large on-campus courses use one or more of these platforms to motivate interaction among peers and between students and instructors [10]. Gelman et al. investigated the emergence of interest-based subcultures in online communities and how they engage a large number of learners [5]. Sajjadi et al. adopt a peer-grading mechanism in a MOOC to explore the effective metric of aggregating peer assessments [24]. Krishnan et al. developed a self-organizing collective system called the Collective Discovery Engine (CDE) to collect insights from a diverse group on how social media can improve learning [8]. Woolley et al. quantify group performance by a collective intelligence factor and suggest that group performance exceeds individual performance [21]. Despite the various crowdsourcing applications, this approach has not been applied to student evaluations. Typical student evaluations involve anonymous individual responses to the questions where students are not allowed to discuss or view each other's response. This mechanism is inefficient because it requires extensive time from the instructors to read each of the responses, and the responses contain many repetitions. M-CAFE 2.0 aims to overcome these challenges by adopting a collaborative filtering mechanism, where students provide peer-to-peer ratings on suggestions with the interested topic and assign a reputation score to each suggestion. The set of suggestions with the highest reputation are presented to the instructors each week.

2.3 M-CAFE 2.0 INTERFACE SUMMARY AND CHANGES

After entering M-CAFE 2.0 (Figure 1a), students specify if they are new users or returning users. New users are prompted to register using their email and are encouraged to provide demographics information, such as gender, age, home country and their primary reason for taking the course (Figure 1b). Then they rate five quantitative assessment topics (QAT) on each separate page on a scale of 0 to 10: Course Difficulty, Course Usefulness, Self-enthusiasm, Self-performance and Homework Effectiveness (Figure 1c). Next, students are shown the comparative plots of their past ratings and the course average ratings on the QATs (Figure 1d). Then students enter the discussion space about the topic "In what specific way would you improve this course this semester or in the future?" (Figure 2). Students click on the tagged spheres to view the suggestions previously provided by their peers on the specific topic tag, evaluate the suggestions on a scale of 0-10 (Figure 1e) and suggest new suggestions with the appropriate topic tag (Figure 1f).

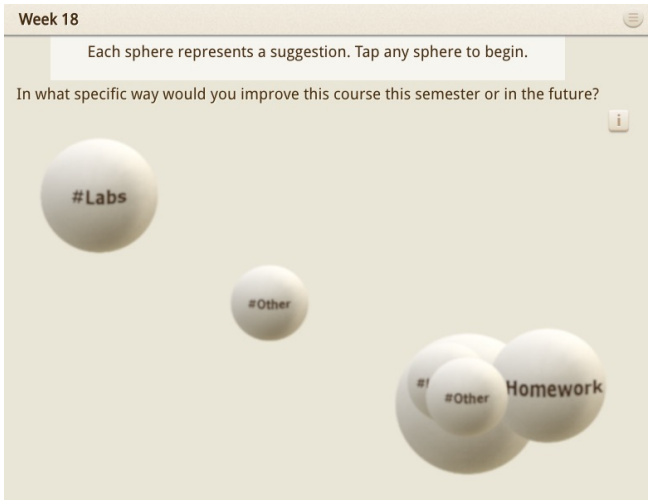

Figure 2: the space of course improvement suggestions with topic tags in M-CAFE 2.0.

Compared with M-CAFE 1.0, the new version, M-CAFE 2.0, adopts a more intuitive interface with a keypad design for ratings. In addition, M-CAFE 2.0 introduces comparative plots displaying the rating history of the current user against the course weekly average for each QAT, enabling users to track their own performance and quickly compare themselves to the course average (Figure 1d). As show in Figure 2, M-CAFE 2.0 further deploys topic tagging in the discussion phase to bring more structure to the textual suggestions. The list of tags includes Exams, Homework, Labs, Lectures, New Topic, Logistics, Projects and Other.

3. PILOT STUDIES AND ANALYSIS

We used M-CAFE 2.0 in two UC Berkeley undergraduate in-person courses to investigate its effectiveness: IEOR 115: Commercial Database Systems in Fall 2015 (F15) and IEOR 170: Industrial Design and Human Factor in Spring 2016 (S16). We compare the results of IEOR 170 in Spring 2015 (S15) with M-CAFE 1.0 usage and IEOR 115 in Fall 2016 (F16) without M-CAFE usage to illustrate that absolute comparison of numerical ratings are unreliable and adopting M-CAFE 2.0 encourages a more diverse set of course improvement suggestions. All four courses were taught by Prof. Ken Goldberg. The two IEOR 170 courses covered the same set of topics/materials on industrial design concepts with similar weekly progress. The two IEOR 115 courses were also similar in content and introduced industrial and commercial database systems. Approximately 60 students enrolled in each course and students in IEOR 170, F16 and IEOR 115, F15 were encouraged to participate anonymously in M-CAFE 2.0 at the beginning of the semesters. Participation in M-CAFE 2.0 is fully voluntary, thus the results suffer from self-selection bias. Furthermore, for IEOR 170, S15, the instructor adopted M-CAFE 1.0 and frequently discussed M-CAFE feedback in class and reminded students to continue participating on a weekly basis. However, for IEOR 115, F15 and IEOR 170, S16, the instructor put less emphasis on M-CAFE 2.0, resulting in a lower participation rate.

Table 1: Participation statistics in different stages of M-CAFE for IEOR 115, F2015 and IEOR 170, F15 and F16, and participation of traditional mid-term evaluation in IEOR 115, F16

	IEOR 170, S15	IEOR 115, F15	IEOR 170, S16	IEOR 115, F16
M-CAFE version	1.0	2.0	2.0	Not used
Total user count	58	57	54	36
Mean weekly user count	32	32	17	N/A
QAT set rating count	474	483	254	N/A
Suggestion count	270	110	90	34
Peer-to-peer rating count	2,483	1,759	979	N/A
Date range of the course	Jan - May, 2015	Sep - Dec, 2015	Jan - May, 2016	Sep – Dec, 2016
Term length of the course	15-week	15-week	15-week	15-week

3.1 Quantitative Evaluation Consistency

Comparing M-CAFE course ratings of two IEOR 170 courses offered in S15 and S16, we find that absolute course assessment rating is not reliable. Figure 3 shows the mean weekly ratings of Course Difficulty from IEOR 170 in 2015 and 2016. Both trends increase gradually as the semester proceeds. Interestingly, there exists a gap between the two ratings from different years throughout the semester. Since the two courses were nearly identical in terms of instructor, material and schedule, there is no reason to believe that the course offered in 2016 was significantly harder than the same course offered in 2015. We suspect that the different rating scales of participating students in the two courses led to this gap, confirming that absolute course evaluation ratings can be unreliable when the population changes. For example, for some course material, one student may assign a Course Difficulty score of 6 but another student may assign a difficulty score of 8. Furthermore, displaying rating history to students provides a benchmark rating scale, which reinforces the difference between the two course ratings in later weeks. Thus we see a consistent relative change over time in average Course Difficulty ratings but a statistically significant difference in absolute ratings between the two courses. Comparing the weekly ratings from the two courses using a t-test, we see

that with 5% significance level, the difference in means each week in the two years is not equal to 0, with an overall mean of 3.57 in 2015 and 5.845 in 2016. Similar gaps are found in the other 4 QATs between the two courses.

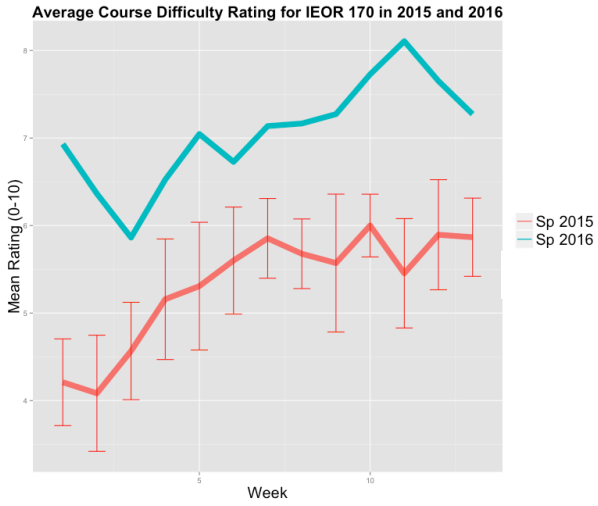

Figure 3: blue line: mean Course Difficulty ratings for 15 weeks in IEOR 170, S16; red line: mean and 2 standard errors above and below the average Course Difficulty ratings for 15 weeks in IEOR 170, S15.

3.2 Qualitative Evaluation Structure and Diversity

Two major challenges in the qualitative part of end-of-course evaluations are the analysis difficulty of unstructured data and the limitation of comment variety resulting from repetition. M-CAFE 2.0 addresses these challenges by collecting textual suggestions from students using topic tagging. For instance, after articulating a suggestion, students are required to choose the appropriate topic tag from a dropdown containing {Exams, Homework, Labs, Lectures, Logistics, New Topics, Projects, Other}. If the student chooses "Other," he/she is encouraged to suggest a new topic tag. For the following analysis, we define "Course Topic" as the topic tags associated with the suggestions, i.e., Exams, Homework, etc. and "Course Module" as the topic of the course materials, i.e., SQL, Relational Schema, etc.

M-CAFE 2.0 organizes feedback by topic tags and encourages students to evaluate the course on a weekly basis when different course modules are covered.

Initial results suggest that:

(1) Students are more likely to provide a new suggestion for topics they have not considered in the peer-rating phase.

By requiring students to provide at least two peer ratings before supplying their own, M-CAFE 2.0 aims to reduce suggestion repetition. After investigating data from the two courses that used M-CAFE 2.0, we find that in both courses more than half of the students supplied a new suggestion with a topic different from the topics that they rated in the peer-rating phase, indicating an intention to articulate new suggestions that are different from those already in the system. Eighty percent of the students in IEOR 115 and 76% of students in IEOR 170 articulated a suggestion from a topic different from the topic of their last rated suggestion. For the students who rated at least one suggestion of the same topic, the content of the new suggestion is different from the rated suggestion. For example, a student in IEOR 115 rated 6 suggestions before providing his/her own, 3 on lectures, 1 on projects, 1 on homework and 1 on other. The suggestions on lectures he/she rated are:

1. "When drawing E-R Diagrams on the board--or any other diagrams that are complex--plan it out so that none of it has to be erased/moved to another board. It is way harder to fix diagrams on paper as we take notes."
2. "Slow down a little bit."
3. "The discussion sections could be clearer. It is difficult to see what kind of table manipulations are going on."

And the student provided the following new suggestion on lectures:

"If possible, posting an outline of every lecture will be very helpful considering the fast pace of the lecture. As a result, we can pay more attention to the explanation instead of putting too much effort in copying down everything in the notes."

Below is another example of a student who rated two suggestions on Homework and further provided a suggestion on Homework. The two suggestions he/she rated:

1. "I hope we receive plenty of feedback on what to improve from graders on our homework and exams."
2. "Re-evaluate homework length. The current homework on designing and prototyping a restraint was too much to ask for in one week -- the design and report alone took me 5 hours. Prototyping was much harder this week due to reduced Jacobs access hours, and my friends and I were not able to make a physical prototype. This homework in particular should have its point balance changed to not take a physical prototype into account."

And the suggestion he/she provided afterwards:

"The fusion360 tutorial in class is not helpful for a first time user at all, and the assignment this week is not easy for students who never use 3D graphic software before."

In case of this student, we suspect that this student is interested in the topic "Homework". Thus he/she purposefully clicked on the spheres with the Homework tag to see if any other students have already suggested the same suggestion. After finding that these two existing suggestions are different from his/hers, he/she articulated his own.

(2) Course improvement suggestions from M-CAFE 2.0 are more diverse than those from traditional mid-course evaluations.

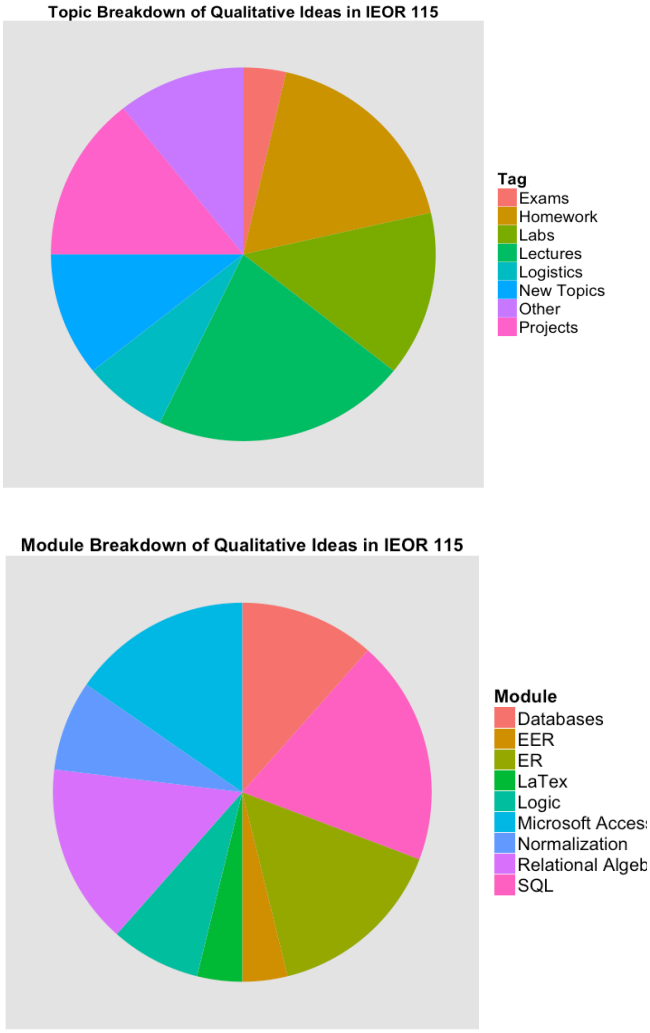

Figure 4: (a) Pie plot of IEOR 115 in 2015 indicating the breakdown of qualitative suggestions in terms of course topics; (b) Pie plot of IEOR 115 in 2015 indicating the breakdown of qualitative suggestions in terms of course modules.

We compare two IEOR 115 courses offered in 2015 and 2016 to observe their suggestion diversity. IEOR 115, F15 used M-CAFE 2.0 and encouraged students to provide qualitative suggestions on a weekly basis, whereas the same course with the same instructor, materials and schedule offered in 2016 didn't use M-CAFE 2.0. Instead, a paper-based mid-term unofficial course evaluation was conducted in lecture on Oct. 3, 2016. We compare the paper results to the qualitative suggestions on M-CAFE 2.0 suggested before Oct. 3, 2015. For IEOR 115, F15, M-CAFE 2.0 collected a total of 50 unique suggestions with topic distribution shown in the pie plot (Figure 4 a), with 13 on Homework, 5 on Labs, 16 on Lectures, 2 on Logistics, 4 on New Topics, 1 on Policies, 5 on Projects and 4 on Other, covering most topics of the course. Out of the 50 suggestions, 35 suggestions mention a specific course module. Figure 4b displays the proportion of suggestions on each course module and results show that most modules received improvement suggestions and the number of suggestions per module is positively correlated with the number of lectures the instructor spent on the module. The suggestions range from "I hope we receive plenty of feedback on what to improve from graders on our homework and exams." to "Despite the pace of the lecture, the examples given in the lecture so far are helpful to understand the overall concept of entity-relationship diagram. For the future, I think it would be better if the Professor can elaborate more why he does a particular step."

For IEOR 115, F16, we manually analyzed the mid-term evaluations from 36 students and summarized the textual suggestions. Out of all the suggestions, 18 are unique within the list, covering multiple course aspects such as Lectures, Homework, Labs, Logistics, etc. However, many students provided the same suggestion: for example, 5 students requested the instructor "to provide a study guide/notes for the lectures." Seventeen of the 18 suggestions are general statements that do not refer to any particular course module, thus making it impossible for the instructor to understand how the students perceive each course module.

4. CONCLUSION

M-CAFE 2.0 collects ongoing student course evaluations and effectively identifies valuable suggestions. Two pilot studies suggest that visualizing relative changes in course assessment over time and topic tagging encourages a more diverse set of course suggestions.

5. FUTURE WORK

This paper provides an evaluation of the effectiveness of M-CAFE 2.0 in regular on-campus courses. In the future, we would like to evaluate the performance of M-CAFE 2.0 in MOOCs/online courses and conduct comparative analysis between the two types of course.

ACKNOWLEDGEMENTS

This study is IRB approved by UC Berkeley under CPHS Protocol 2014-04-6297. This work is supported in part by the Blum Center for Developing Economies and the Development Impact Lab (USAID Cooperative Agreement AID-OAA-A-12- 00011) as part of the USAID Higher Education Solutions Network, UC Berkeley's Algorithms, Machines, and People (AMP) Lab, and the Connected Communities Initiative at the Center for Information Technology Research in the Interest of Society (CITRIS) and the Banatao Institute at UC Berkeley. This work is also supported in part by the Philippine Commission on Higher Education through the Philippine-California Advanced

Research Institutes Project (no. IIID-2015-07). Thanks to Animesh Garg, Allen Huang, Allison Cliff and Steve McKinley. Thanks to all students who participated and colleagues and reviewers who provided feedback on earlier drafts.

REFERENCES

1. Adams, M.J. and Umbach, P.D., 2012. Nonresponse and online student evaluations of teaching: Understanding the influence of salience, fatigue, and academic environments. *Research in Higher Education*, *53*(5), pp.576-591.
2. Daradoumis, T., Bassi, R., Xhafa, F. and Caballé, S., 2013, October. A review on massive e-learning (MOOC) design, delivery and assessment. In *P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), 2013 Eighth International Conference on* (pp. 208-213). IEEE.
3. Flaherty, C. (2016). Zero Correlation Between Evaluations and Learning. Retrieved from https://www.insidehighered.com/news/2016/09/21/new-study-could-be-another-nail-coffin-validity-student-evaluations-teaching.
4. Freeman, R. and Dobbins, K., 2013. Are we serious about enhancing courses? Using the principles of assessment for learning to enhance course evaluation. *Assessment & Evaluation in Higher Education*, *38*(2), pp.142-151.
5. Gelman, B.U., Beckley, C., Johri, A., Domeniconi, C. and Yang, S., 2016, April. Online Urbanism: Interest-based Subcultures as Drivers of Informal Learning in an Online Community. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale* (pp. 21-30). ACM.
6. Greene, J.A., Oswald, C.A. and Pomerantz, J., 2015. Predictors of retention and achievement in a massive open online course. *American Educational Research Journal*, p.0002831215584621.
7. Khong, T.L., 2016. The Validity and Reliability of the Student Evaluation of Teaching: A case in a Private Higher Educational Institution in Malaysia.*International Journal for Innovation Education and Research*, *2*(9).
8. Krishnan, S., Patel, J., Franklin, M. and Goldberg, K., 2014. Social influence bias in recommender systems: a methodology for learning, analyzing, and mitigating bias in ratings. In *Proceedings of the 8th ACM Conference on Recommender systems* (pp. 137-144).
9. Krishnan, S., Okubo, Y., Uchino, K. and Goldberg, K., 2013. Using a social media platform to explore how social media can enhance primary and secondary learning. In *Learning International Networks Consortium (LINC) 2013 Conference*.
10. Mak, S., Williams, R., & Mackness, J. (2010). Blogs and Forums as Communication and Learning Tools in a MOOC. In *Proceedings of the 7th International Conference on Networked Learning*, ISBN 978-1-86220-225-2, p.275-284.
11. Marsh, H.W., and Roche, L.A. Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist*, 52(11). (1997)
12. McCullough, B.D. and Radson, D., 2011. Analysing student evaluations of teaching: comparing means and proportions. *Evaluation & Research in Education*, *24*(3), pp.183-202.
13. Morley, D., 2014. Assessing the reliability of student evaluations of teaching: choosing the right coefficient. *Assessment & Evaluation in Higher Education*,*39*(2), pp.127-139.
14. Pursel, B.K., Zhang, L., Jablokow, K.W., Choi, G.W. and Velegol, D., 2016. Understanding MOOC students: motivations and behaviours indicative of MOOC completion. *Journal of Computer Assisted Learning*, *32*(3), pp.202-217.
15. Sajjadi, M.S., Alamgir, M. and von Luxburg, U., 2016, April. Peer Grading in a Course on Algorithms and Data Structures: Machine Learning Algorithms do not Improve over Simple Baselines. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale* (pp. 369-378). ACM.
16. Sandeen, C., 2013. Assessment's place in the new MOOC world. *Research & practice in assessment*, *8*.
17. Spooren, P., Brockx, B. and Mortelmans, D., 2013. On the validity of student evaluation of teaching the state of the art. *Review of Educational Research*,*83*(4), pp.598-642.
18. Stark, P.B. and Freishtat, R., 2014. An evaluation of course evaluations.*Center for Teaching and Learning, University of California, Berkley*.
19. Surgenor, P.W.G., 2013. Obstacles and opportunities: addressing the growing pains of summative student evaluation of teaching. *Assessment & Evaluation in Higher Education*, *38*(3), pp.363-376.
20. Wang, Y., 2013, June. Exploring possible reasons behind low student retention rates of massive online open courses: A comparative case study from a social cognitive perspective. In *AIED 2013 Workshops Proceedings Volume* (p. 58).
21. Woolley, A.W., Chabris, C.F., Pentland, A., Hashmi, N. and Malone, T.W., 2010. Evidence for a collective intelligence factor in the performance of human groups. *science*, *330*(6004), pp.686-688.
22. Xiong, Y., Li, H., Kornhaber, M.L., Suen, H.K., Pursel, B. and Goins, D.D., 2015. Examining the Relations among Student Motivation, Engagement, and Retention in a MOOC: A Structural Equation Modeling Approach. *Global Education Review*, *2*(3).
23. Zheng, S., Rosson, M.B., Shih, P.C. and Carroll, J.M., 2015, February. Understanding student motivation, behaviors and perceptions in MOOCs. In*Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 1882-1895). ACM.
24. Zhou, M., Cliff, A., Krishnan, S., Nonnecke, B., Crittenden, C., Uchino, K., & Goldberg, K. 2015, September. M-CAFE 1.0: Motivating and Prioritizing Ongoing Student Feedback During MOOCs and Large on-Campus Courses using Collaborative Filtering. In *Proceedings of the 16th Annual Conference on Information Technology Education* (pp. 153-158). ACM.