Re-designing the Structure of Online Courses to Empower Educational Data Mining

Zhongzhou Chen
University of Central Florida
4111 Libra Drive
PSB 153
+1 321-236-8568
Zhongzhou.Chen@ucf.edu

Sunbok Lee
University of Houston
3695 Cullen Boulevard
Heyne building 231D
+1 7816459203
slee95@Central.UH.edu

Geoffrey Garrido
University of Central Florida
4111 Libra Drive
PSB 153
+1 3219873375
Geoff.garrido@knights.ucf.edu

ABSTRACT

The amount of information contained in any educational data set is fundamentally constrained by the instructional conditions under which the data are collected. In this study, we show that by re-designing the structure of traditional online courses, we can improve the ability of educational data mining to provide useful information for instructors. This new design, referred to as Online Learning Modules, blends frequent learning assessment as seen in intelligent tutoring systems into the structure of conventional online courses, allowing learning behavior data and learning outcome data to be collected from the same learning module. By applying relatively straightforward clustering analysis to data collected from a sequence of four modules, we are able to gain insight on whether students are spending enough time studying and on the effectiveness of the instructional materials, two questions most instructors ask each day.

Keywords

Online Instructional Design; Clustering Analysis; Data Interpretability; Supporting Teachers

1. INTRODUCTION

The central goal of educational data mining is to “mine educational data sets to answer educational research questions that shed light on the learning process”. To this end, the predominant focus of the EDM community has been on developing and advancing methods and algorithms to effectively extract information from existing educational data sets. However, the amount of information contained in any given data set is fundamentally constrained by the instructional conditions under which the data is collected [17], such as the nature of the learning tasks, the design and organization of instructional contents, and even the available features of the educational platform. As a simple example, if the final exam is the only assessment administered in an online course, then information about students’ content mastery at any other time during the course is obviously not contained in the data. Therefore, we ask the question: is it possible to enhance the ability of EDM to provide useful information for instructors, by re-designing the structure of the online course to improve the quality of the data that it produces?

Many of today’s online courses more or less inherited their structure from their off-line, face-to-face predecessors. For example, many MOOCs are created directly based on existing face-to-face courses [9, 29, 33]. Those courses typically contain a variety of learning resources, from e-text and videos to problems and forums, organized into week-long units. This structure allows students to display a plethora of different learning behaviors, which has become the focus of many recent studies in EDM. [2, 14, 20, 24, 27]

On the other hand, students’ learning outcome is assessed relatively sparsely in a typical online course. Many recent studies still use “certification rate” or “retention rate” as a proxy for learning over the entire course [14, 21, 27], which can be problematic [19]. Moreover, very few online courses contain any form of pre-test [12],. This is particularly problematic for learning measurement in MOOCs, as there are significant variations in students’ incoming knowledge and background [11, 19]. Insufficient assessment of learning outcome made it difficult for researchers to make meaningful correlations between learning behavior and learning outcome.

In contrast, students’ knowledge state is being constantly assessed in intelligent tutoring systems (ITS), another online instructional system widely studied by the EDM community [4, 15, 18, 30]. A number of methods have been developed to measure students’ learning progress in a ITS with high resolution [13, 22, 23]. However, students’ learning behavior is much more restricted in many ITS as compared to online courses, and oftentimes instructional materials in a ITS consist of only simple hints or feedback texts.

Can we re-design the structure of online courses to include certain features of ITS so that it contains more frequent and accurate learning assessment, while still providing enough freedom for students to display a variety of learning behavior? In this paper we present such an attempt at combining the advantages of both systems, by constructing a small online course consisting of a sequence of four Online Learning Modules (OLMs). Each module contains both instruction and assessment, which enables us to make correlated measurements on students’ learning behavior and learning outcome in close proximity. Moreover, students are required to make one attempt on the assessment before accessing the instruction, which serves as a de-facto pre-test for each learning module. We demonstrate that by applying relatively simple data mining algorithms, data produced by OLMs could provide valuable insight on two questions that every instructor encounters on a daily basis: Q1: Are students spending enough time and effort studying the materials? Q2: How effective are the instructional resources in the course?

Both questions are best answered when considering learning behavior and learning outcomes together. For Q1, “enough time” is best defined for a given instructional resource when students spending less than that time have poorer learning outcomes; For Q2, “effectiveness” can be more accurately measured from the learning outcome of students who spent adequate time and effort learning from the resources. In the remainder of this paper, we will first introduce the design of OLMs and implementation of the current study, then describe the data collection procedure, analysis and visualization methods, followed by the outcomes of the study and ending with a discussion of the impact of this study on potential future research.

2. METHODS

2.1 Design of OLMs

The design of OLM is inspired by research on deliberate practice [16] and mastery-based learning [7, 8], and in particular influenced by the design of the ASSISTMENTs tutoring platform [3, 18]. Each OLM module contains an instructional component (IC) and an assessment component (AC) (Figure 1). The IC consists of both instructional text and ungraded practice problems separated into multiple pages, focused on teaching a single physics concept or a problem-solving skill. Students receive immediate feedback and have access to the problem solution after attempting any practice problem. Each IC typically takes about 10 minutes to an hour for a student to finish, which resembles a small unit in an online course. The AC consists of either 2-3 simple multiple-choice concept problems or 1 complex multiple-choice problem, depending on the focus of the module.

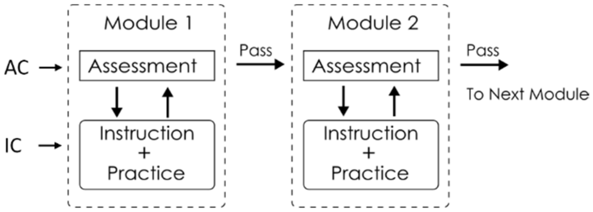

Figure 1: Schematic representation of the structure of OLM and OLM sequence

A series of OLMs are combined sequentially to form a learning unit on a given topic. A student passes a module by correctly answering all the questions in the assessment component, and can proceed onto the next module only after passing the current one. Each student can have multiple attempts on the AC. On each new attempt, a slightly different version of the assessment problem(s) drawn from a problem bank is presented to the student.

A key feature of OLM is that students are required to make at least one attempt on the assessment before being given access to the IC. After the initial attempt, students can either study the IC, or make additional attempts on the assessment. On each attempt the student is presented with a slightly different problem until the problem bank in the assessment component is depleted. During an attempt the IC is temporarily locked from access.

The OLM design has three major advantages for data collection and analysis: First, students’ AC attempts before and after instruction serve as de-facto pre and post-tests, increasing the accuracy and frequency of learning measurement. Second, the length and types of learning resources in the IC allows for a richer variety of student learning behavior to be observed compared to many ITS. Finally, by combining instruction and assessment into one module, it allows for observations of learning outcome and learning behavior to be interpreted in the context of each other.

2.2 Study Design and Data Collection

Individual OLMs were created on the award-winning learning objects platform, Obojobo, developed by the Learning System and Technology (LS&T) team at the Center for Distributed Learning at University of Central Florida [6], and administered to students as a sequence via the Canvas learning management system. For the current study, student subjects were recruited from three sections of calculus-based college introductory physics course at University of Central Florida during the Spring 2017 semester. The OLMs were provided to students as an optional reviewing tool for an upcoming exam.

Four OLMs were created on the topic of conservation of mechanical energy with each module focusing on a single concept or a problem-solving skill. The problem bank of each AC contains 3 isomorphic multiple-choice problems authored based on published assessment instruments in physics[32]. The distractors in each problem are designed to capture common student misconceptions.

The number of students who made at least 1 attempt on the AC of modules 1-4 are 75, 54, 47 and 40 respectively. In this study, students were allowed 50 attempts on each module to ensure that they can all proceed to the next module.

Time-stamp data on the following types of student events are collected by the Obojobo platform: Entering and exiting a page in both IC and AC; Starting and finishing an attempt on either an assessment problem or a practice problem; Viewing a practice or assessment problem; Submitting an answer to a practice or assessment problem; Outcome of each attempt at the AC.

2.3 Data Analysis

2.3.1 Capturing Learning Behavior within Longest Study Session

All of the interactions by one student with the IC that took place between two consecutive assessment attempts are treated as a single “study session” (SS). A student can have multiple SS by going back and forth between the IC and the assessment component. For answering the questions in this manuscript, we only consider SS that took place before the first time a student passes the assessment component is recorded.

In a total of 168 occasions where a student interacted with the IC of a module, 76% (127) of the time all interactions took place in a single SS. In most of the other occasions, there is a major SS that is significantly longer than the other SS. In only 4 cases did the second longest SS reach at least 50% as long as the longest SS (LSS). Since the majority of students’ learning behavior for each module took place during their LSS, it serves as a good approximation for measuring students’ learning effort of the given module. For the current analysis, students’ learning behavior within the LSS is characterized along three dimensions:

1. The duration of the LSS, measured as the sum of the times spent on each accessed page in the IC.

2. The average number of attempts made on practice problems, measured as the total attempts made divided by the number of practice problems viewed by the student.

3. The percentage of contents accessed, measured as the sum of page entering events plus problem viewing events, divided by the sum of the number of pages and the number of practice problems in each module.

2.3.2 *Clustering Analysis of students' learning behavior*

In this study, we assume that students' learning behavior will form multiple clusters due to different learning strategies, habits and incoming knowledge states. In order to identify such subgroup, we used a mixture model in which the whole population distribution is represented by the sum of component distributions representing subgroups, and the probabilities of students' belonging to subgroups or classes are estimated. We used Mplus software [28]to fit the mixture model to our data. The optimal number of classes was judged based on six statistical indices provided by Mplus: Akaike Information Criterion (AIC)[1], Bayesian Information Criterion (BIC)[31], Sample-size Adjusted BIC, Vuong-Lo-Mendell-Rubin Likelihood Ratio (VLMRLR) Test[34], Lo-Mendell-Rubin Adjusted LRT (LMRALRT) test[25], and Bootstrap Likelihood Ratio (BLR) test[26]. AIC, BIC, and sBIC are goodness of fit indices which consist of -2 log (likelihood) and an additional term for penalizing a complex model. Each tries to strike balance between fit (-2 log (likelihood)) and parsimony (a penalty term), and a smaller value indicates better fit. The other three indices are the statistical tests comparing how well the data is fitted by models with n and n-1 classes, i.e. p-value less than .05 from those tests indicates that the current model with n-classes has a significantly better fit than the model with (n-1)-classes. In short, the optimal number of classes can be determined by running mixture models with a different number of classes (e.g., models with 1,2,3, and 4 classes) and by selecting the model showing the overall best fit to the data based on those six indices.

2.3.3 *Categorizing Learning Outcome*

Students' learning outcome from each module can be classified into four classes according to performance in the AC and time of attempt relative to the LSS:

1. **Initial Pass (InitP):** Passing the AC within 2 attempts before LSS. Those students did not need to learn from the IC, although a small fraction still interacted with the IC. An earlier study on students' test-taking effort on the initial attempt estimated that 80-85% of the students took the attempt seriously. [10]

2. **Effective (Eff):** Passing the AC within 2 attempts after LSS (not including attempts before LSS).

3. **Ineffective (Ineff):** Passing the AC using more than 2 attempts after LSS.

4. **Abort:** Never passing the AC, thus cannot access the next module in the sequence.

In addition, in a few cases a student passes the AC using more than 2 attempts without accessing the IC. Since those students are more likely to be randomly guessing the answer rather than actually doing the problem, we also categorized them as "Abort".

3. RESULTS AND DISCUSSION

3.1 Results

3.1.1 *Identifying Clusters of Learning Behavior*

Cluster Analysis was performed on all three dimensions of learning behavior for each module, for all students who didn't pass the module on their attempt before LSS. The 3-dimensional clustering analysis did not converge for any module likely due to small sample size. Clustering analysis on both average number of attempts and percentage of content accessed always favored single cluster for every module. The mean average number of practice problem attempts are between 1 and 3 attempts for all modules, and the mean content accessed is more than 95% for all modules.

For the time-on-task dimension, initial clustering results were significantly distorted by a few data points with extremely long and scattered LSS durations, most likely due to students leaving their computer without logging off of the system or idling. Thus, clusters with less than 5 students and significantly larger mean values were removed and the clustering analysis re-run, until no such cluster existed. We also found a small cluster of students with mean LSS time of 30 seconds and interacted with the IC of Module 1 only. Those students were also removed since they are likely students who are curious about the new system but did not seriously study the content. The resulting statistical indices for different number of clusters are listed in TABLE 1.

TABLE 1: Statistical indices of mixture-model clustering analysis. Favorable values are highlighted in red.

Module	class	AIC	BIC	sBIC	VLMR (p)	LMR (p)	BLRT (p)
Module 1 (N = 36)	1	38.6	41.8	35.5	NA	NA	NA
	2	37.4	45.3	29.7	0.05	0.07	1.00
	3	37.3	50.0	25.0	0.08	0.11	0.43
	4	did not converge					
Module 2 (N = 38)	1	100.4	103.6	96.4	NA	NA	NA
	2	88.6	96.8	81.2	0.01	0.01	0.00
	3	90.8	103.9	78.9	0.56	0.58	1.00
	4	90.0	108.0	73.6	1.00	1.00	1.00
Module 3 (N = 37)	1	119.4	122.7	116.4	NA	NA	NA
	2	116.3	124.3	108.7	0.02	0.03	0.15
	3	116.6	129.5	104.5	0.02	0.03	1.00
	4	116.8	134.5	100.1	0.27	0.30	1.00
Module 4 (N = 26)	1	95.2	97.7	91.5	NA	NA	NA
	2	90.1	96.4	80.9	0.01	0.01	1.00
	3	88.2	98.3	73.4	0.05	0.06	1.00
	4	88.5	102.3	68.2	0.14	0.17	1.00

For all four modules, a 2-cluster model is either most favorable, or equally as favorable as a 3-cluster model. Therefore, we adopt a 2-cluster model for the LSS duration dimension for each module, referring to the cluster with shorter mean time as "Brief" and the longer mean time as "Extensive" (Ext). One possible interpretation is that the "Brief" clusters consist of students who had some level of initial understanding and needed a quick refresh of the content knowledge, while the "Extensive" clusters are students who failed to learn the content properly during regular lecture, and are actually learning from the IC of the modules.

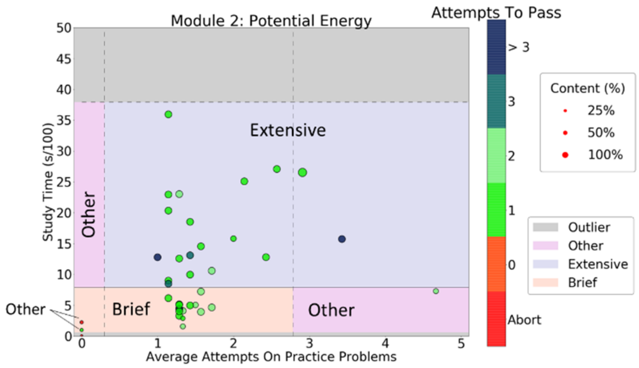

Figure 2: Example of refinement of clustering analysis outcome: Horizontal solid line divides the two clusters. The vertical dashed line indicates 1.5 standard deviation from population mean.

The clusters are further refined by labeling a few students who displayed inconsistent behavior along the other two dimensions as "other". As illustrated in Figure 2 , a student in the "brief" cluster who makes significantly more attempts on practice problem (more than 1.5 sd above the group mean) is labeled as "other" since his/her learning behavior is very different from other students (lower right purple area). Similarly, any student in the "Extensive" cluster whose average practice problem attempt or percentage of content accessed is 1.5 sd less than the population mean is also labeled as "other", since the student is most likely not meaningfully engaged with the learning material. Finally, a few students who didn't attempt any practice problems, and/or interacted with the IC for less than 60 seconds are also labeled as "other" as their learning behavior is significantly different from the rest of the population.

The refinement strategy is illustrated in Figure 2 using data from Module 2 as an example. The final clusters of students' learning behavior are listed in TABLE 2.

TABLE 2: Refined learning behavior clusters

Modules	Cluster	N	mean (s)	Var. (s)
1	Brief	25	519	75
	Ext	8	1272	12
	Other	3	NA	NA
2	Brief	15	416	36
	Ext	19	1594	675
	Other	4	NA	NA
3	Brief	18	1233	495
	Ext	16	3382	165
	Other	3	NA	NA
4	Brief	18	1141	576
	Ext	5	4175	256
	Other	3	NA	NA

3.1.2 Combining Learning Behavior with Learning Outcome

To visually represent the relation between learning behavior and learning outcome in each module, we plot both types of information together in four sunburst charts shown in Figure 3. The inner rings show the distribution of four classes of learning outcome, while the outer ring shows the distribution of the three learning behavior clusters within each learning outcome classes. Some of the key observations from the data are summarized in TABLE 3.

Looking at assessment performance alone, Modules 3 and 4 are significantly harder than modules 1 and 2, judging by both the fraction of students in InitP (Fisher's exact test, $p = 0.01$) and the total fraction of students who passed the module either before or after accessing the IC (Tot.Pass) ($p < 0.01, \chi^2 = 40, df = 3$). The total number of passing students is the sum of the InitP group and the Eff group.

A noteworthy observation is that initially less students passed module 2 than module 1, but after studying the IC the trend was reversed.

The effectiveness of the IC can be estimated by the ratio of the size of Eff vs. Ineff classes. For simplicity, in the remainder of the paper (including TABLE 3) we will include the students in the "Abort" class into the "Ineff" class, which now contain all students who failed to pass within two attempts after LSS. Modules 1 and 2 have a significantly higher ratio of Eff vs. Ineff ($p < 0.01, \chi^2 = 34.39, df = 3$). (Test still significant when either module 2 or module 4 is excluded).

From the learning behavior perspective, Modules 2 and 3 have significantly higher Ext vs. Brief ratio ($p = 0.01, \chi^2 = 11, df = 3$) as compared to the other two modules. Somewhat unexpectedly, the size of "Extensive" group in module 4 is the smallest of the four, consisting of only 5 students.

TABLE 3: Main observations. The total number includes students who passed the AC before studying IC

Modules	N	InitP	Tot. Pass	Eff/Ineff	Ext/Brief
1	47	0.26	0.79	2.50	0.32
2	40	0.12	0.88	6.00	1.27
3	35	0.03	0.57	1.27	0.89
4	25	0.04	0.16	0.14	0.28

Finally, the correlation between the learning behavior clusters ("Brief", "Extensive") and the learning outcome measures ("Eff", "Ineff") are not significant when the four modules are combined (Fisher's exact test, $p = 0.35$, $\mathrm{OR} = 0.65$). This correlation is also not statistically significant at $p = 0.05$ level when each of the four modules were tested individually. In other words, there is no significant difference in the probability of passing each module after learning from the IC between the "Brief" and "Extensive" groups.

Of the 61 students that are not excluded as an outlier in at least one of the modules, only 4 are Brief and Ineff (including Abort) for 2 modules, and no student is both Brief and Ineffective for more than 2 modules. In comparison, 3 students are Extensive and Effective for 3 out of 4 modules.

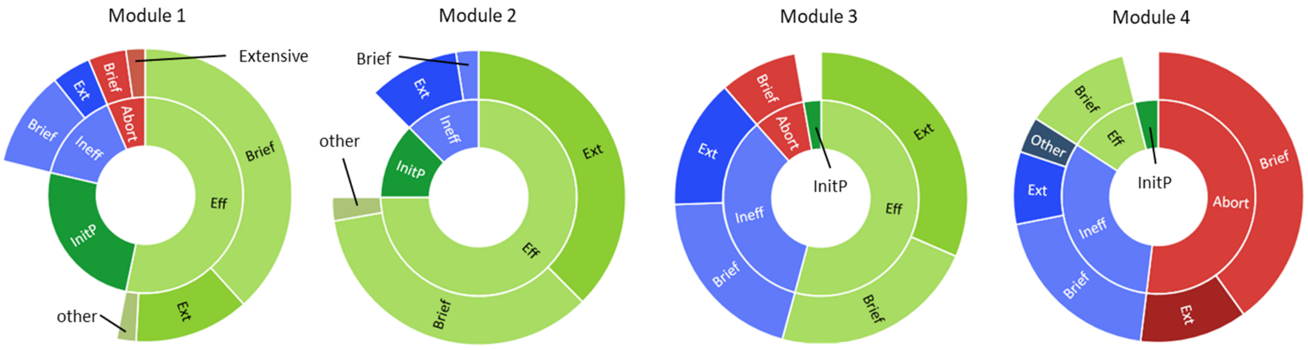

Figure 3: Sunburst charts representing students' learning behavior and learning outcome

3.2 Discussion

By combining learning behavior measurement with learning outcome measurement, we are able to answer both research questions introduced in Section 1 and provide useful information for instructors regarding the four OLMs. For RQ1, data suggests that students in this study are consciously adjusting their learning effort according to their own learning needs and the difficulty of the task. This claim is supported by the lack of correlation between the two learning effort clusters and the three learning outcome clusters, together with the fact that only a few students were consistently "Brief and Ineffective" or "Brief and Abort". In other words, all of the students can be viewed as spending "enough time" on the IC, as there are no clear benefits associated spending longer time. At least, the instructor should be advised to only give the suggestion of "study harder" to the 4 students who are "Brief" and "Ineffective" for 2 out of 4 modules. Had we only considered behavior measurement alone, many more students who have better incoming knowledge on the topic would have been misclassified as less motivated.

One possible explanation for this observation is that since this is a voluntary, not-for-credit activity, only motivated students attempted the OLMs. In future studies it will be interesting to see if the outcome changes when OLMs are being assigned for credit to the entire class.

Our data analysis also provides rich information with regard to the quality of learning resources in the OLMs (RQ2). Among the four modules, Module 1 is the easiest, with high initial passing rate and low "Extensive" vs. "Brief" ratio, suggesting that many students only needed a quick "refresh" of the content. The assessment of Module 2 is slightly harder (lower fraction of InitP), but most students were able to successfully learn the content by carefully studying the IC, as indicated by significantly higher Eff to Ineff ratio and the highest Extensive to Brief ratio. These data suggest that the resources in the IC of Module 2 are effective for the current student population. Note that if only a posttest were given in this course, we might have concluded that problems in modules 2 were easier than those in module 1 without considering students' prior knowledge and learning effort. The AC of Module 3 is even harder, and despite a significant fraction of students in the "Extensive" cluster, a smaller fraction of students passed the AC after studying the IC, suggesting that the instructional resources in the IC of module 3 are less effective and need more improvement.

Module 4 has an unusually large fraction of "Abort" students, and a surprisingly small "Ext" cluster despite being the hardest of all four modules. A likely explanation is that many weaker students find this module too challenging, and lack both the confidence and the incentive to study it as it is the last module in the sequence. In fact, half of the students (9 out of 16) belonging to the "Ext" cluster in Module 3 aborted module 4, whereas only a third (6 out of 18) of students in the "Brief" cluster of Module 3 aborted Module 4.

The majority of the above information is intuitively represented in the sunburst charts in Figure 3, which clearly signals to the instructor that Modules 3 and 4 needs to be improved, and that at least on Module 3, students' lower performance is not caused by insufficient learning effort, but rather ineffective instructional resources.

It is worth pointing out that the mean duration of the "Brief" cluster for modules 3 and 4 are similar to that of the "Ext" cluster for Modules 1 and 2. One possibility is that the learning behavior of the "Brief" cluster in Modules 3 and 4 are more similar to the "Ext" cluster of Modules 1 and 2. However, we only found 4 students who changed from the "Ext" cluster in Module 2 to the "Brief" cluster in Module 3. We think that a more dominant factor is simply that the IC in Modules 3 and 4 contains instructional resources that took longer to go through than Modules 1 and 2. However, examining whether the same cluster across different modules originate from similar learning behavior is an important question for future research.

Finally, we would like to address a couple of detailed choices in both study design and data analysis. First of all, the choice of using 2 attempts instead of one as the threshold for passing a unit is to mitigate the effect of carelessness in students and the possibility of accidentally selecting the wrong choice item. Furthermore, research on multiple attempts has shown that subsequent attempts on problems have equal discrimination power as the initial attempt [5].

Secondly, even though students have already been exposed to the content in lecture, it is clear from the analysis that most of them still need to either refresh or learn the content from the OLMs. We believe that the methods developed in this research are general to most online-courses, especially when we are facing an increasingly diverse student population in higher education and MOOCs in particular.

Finally, choosing mixture-model clustering analysis to capture patterns in students' learning behavior has two major advantages. First, it provides a systematic method to remove outliers in the data, and second, it accommodates the fact that different resources intrinsically require different amounts of time to study, by providing natural cutoffs between "Brief" and "Extensive" clusters.

4. SUMMARY

In this paper, we presented a case where a re-design of the online course structure enabled new methods of data analysis and visualization that provide useful information for instructors. The OLMs are designed to measure both learning behavior and learning outcome in the same module, greatly improving the interpretability of both types of data. Future larger scale studies involving more advanced data mining methods will likely provide insight into even more aspects of students' learning process, such as knowledge transfer, motivation, and meta-cognitive skills.

As data collection and analysis becomes an increasingly important and integrated part of today's technology enhanced education system, it is valuable for data scientists to be more actively involved in the design of instructional systems, resources and environments, rather than simply being on the receiving end of educational data. Design choices that are made to improve the quality of data, even as small as requiring an extra click to view a given problem, may significantly enhance the power of educational data mining, which eventually benefits teaching and learning.

5. ACKNOWLEDGMENTS

We thank the UCF LS&T team led by Dr. Francisca Yonekura for developing the Obojobo platform for implementing OLMs and Dr. Patsy Moskal at the Center for Distributed Learning at UCF for commenting on the manuscript. The project was supported by star-up funds from the University of Central Florida.

6. REFERENCES

[1] Akaike, H. 1974. A New Look at the Statistical Model Identification. 215–222.

[2] An, T.-S. et al. 2017. Can typical behaviors identified in MOOCs be discovered in other courses? *Proceedings of the 10th International Conference on Educational Data*

Mining (2017), 220–225.

[3] Baker, R.S. 2016. Stupid Tutoring Systems, Intelligent Humans. *International Journal of Artificial Intelligence in Education*. 26, 2 (2016), 600–614. DOI:https://doi.org/10.1007/s40593-016-0105-0.

[4] Baker, R.S.J.D. 2010. Data mining for education. *International Encyclopedia of Education*. 7, (2010), 112–118. DOI:https://doi.org/10.4018/978-1-59140-557-3.

[5] Bergner, Y. et al. 2015. Estimation of ability from homework items when there are missing and/or multiple attempts. *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge - LAK '15*. (2015), 118–125. DOI:https://doi.org/10.1145/2723576.2723582.

[6] Bishop, C. et al. 2013. Pilot Study Examining Student Learning Gains Using Online Information Literacy Modules. *Proceedings of the Association of College and Research Libraries (ACRL) Annual Conference* (Indianapolis, Indiatna, 2013), 466–471.

[7] Block, J. and Burns, R. 1976. Mastery learning. *American Educational Research Journal*. 4, (1976), 3–49.

[8] Bloom, B.S. 1968. Learning for Mastery. *UCLA Evaluation Comment*. 1, 2 (1968).

[9] Breslow, L. et al. 2013. Studying learning in the worldwide classroom: Research into edX's first MOOC. *Research & Practice in Assessment*. 8, March 2012 (2013), 13–25. DOI:https://doi.org/10.1007/BF01173772.

[10] Chen, Z. et al. 2018. Designing online learning modules to conduct pre- and post-testing at high frequency. *2017 Physics Education Research Conference Proceedings* (Cincinnati, OH, Jan. 2018), 84–87.

[11] Chen, Z. et al. 2016. Researching for better instructional methods using AB experiments in MOOCs:Results and Challenges. *Research and Practice in Technology Enhanced Learning*. 11, 9 (Dec. 2016). DOI:https://doi.org/10.1186/s41039-016-0034-4.

[12] Chudzicki, C. et al. 2015. Validating the pre/post-test in a MOOC environment. *2015 Physics Education Research Conference Proceedings*. (2015), 83–86. DOI:https://doi.org/10.1119/perc.2015.pr.016.

[13] Cook, J. et al. 2017. Task and Timing: Separating Procedural and Tactical Knowledge in Student Models. *Proceedings of the 10th International Conference on Educational Data Mining* (2017), 186–191.

[14] Davis, D. et al. 2016. Gauging MOOC Learners' Adherence to the Designed Learning Path. *Proceedings of the 9th International Conference on Educational Data Mining*. (2016), 54–61.

[15] Doroudi, S. et al. 2016. Sequence matters, but how exactly? Towards a workflow for evaluating activity sequences from data. *Proceedings of the 9th International Conference on Educational Data Mining* (2016), 70–77.

[16] Ericsson, K.A. et al. 1993. The Role of Deliberate Practice in the Acquisition of Expert Performance. *Psychological Review*. 100, 3 (1993), 363–406.

[17] Gašević, D. et al. 2015. Let ' s not forget : Learning analytics are about learning. *TechTrends*. 59, 1 (2015), 64–71. DOI:https://doi.org/10.1007/s11528-014-0822-x.

[18] Heffernan, N.T. and Heffernan, C.L. 2014. The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*. 24, 4 (2014), 470–497. DOI:https://doi.org/10.1007/s40593-014-0024-x.

[19] Ho, A.D. et al. 2014. HarvardX and MITx: The First Year of Open Online Courses, Fall 2012-Summer 2013. *SSRN Electronic Journal*. 1 (2014), 1–33. DOI:https://doi.org/10.2139/ssrn.2381263.

[20] Kizilcec, R.F. et al. 2013. Deconstructing Disengagement : Analyzing Learner Subpopulations in Massive Open Online Courses. *Lak '13*. (2013), 10. DOI:https://doi.org/10.1145/2460296.2460330.

[21] Li, Y. et al. 2017. When and who at risk ? Call back at these critical points. *Proceedings of the 10th International Conference on Educational Data Mining* (2017), 168–173.

[22] Liu, R. and Koedinger, K.R. 2017. Towards reliable and valid measurement of individualized student parameters. *Proceedings of the 10th International Conference on Educational Data Mining*. (2017), 135–142.

[23] Liu, R. and Koedinger, K.R. 2015. Variations in learning rate : Student classification based on systematic residual error patterns across practice opportunities. *8th International Conference on Educational Data Mining* (2015).

[24] Liu, Z. et al. 2016. MOOC Learner Behaviors by Country and Culture; an Exploratory Analysis. *Proceedings of the 9th International Conference on Educational Data Mining* (2016), 127–134.

[25] Lo, Y. et al. Testing the Number of Components in a Normal Mixture. *Biometrika*. Oxford University PressBiometrika Trust.

[26] McLachlan, G.J. 1987. On Bootstrapping the Likelihood Ratio Test Stastistic for the Number of Components in a Normal Mixture. *Applied Statistics*. 36, 3 (1987), 318. DOI:https://doi.org/10.2307/2347790.

[27] Miyamoto, Y.R. et al. 2015. Beyond time-on-task: The relationship between spaced study and certification in MOOCs. *Journal of Learning Analytics*. 2, 2 (Dec. 2015), 47–69. DOI:https://doi.org/10.18608/jla.2015.22.5.

[28] Muthén, L.K. and Muthén, B.O. *Mplus User's Guide Seventh Ediction*. Muthén & Muthén.

[29] Rayyan, S. et al. 2016. A MOOC based on blended pedagogy. *Journal of Computer Assisted Learning*. 32, 3 (2016), 190–201. DOI:https://doi.org/10.1111/jcal.12126.

[30] Romero, C. and Ventura, S. 2010. Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*. 40, 6 (2010), 601–618. DOI:https://doi.org/10.1109/TSMCC.2010.2053532.

[31] Schwarz, G. Estimating the Dimension of a Model. *The Annals of Statistics*. Institute of Mathematical Statistics.

[32] Singh, C. and Rosengrant, D. 2003. Multiple-choice test of energy and momentum concepts. *American Journal of Physics*. 71, 6 (2003), 607.

DOI:https://doi.org/10.1119/1.1571832.

[33] Toven-Lindsey, B. et al. 2015. Virtually unlimited classrooms: Pedagogical practices in massive open online courses. *Internet and Higher Education*. 24, (2015), 1–12. DOI:https://doi.org/10.1016/j.iheduc.2014.07.001.

[34] Vuong, Q.H. 1989. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica*. 57, 2 (Mar. 1989), 307. DOI:https://doi.org/10.2307/1912557.