Computers & Education 75 (2014) 282–292

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Modelling and quantifying the behaviours of students in lecture capture environments

Christopher Brooks*, Graham Erickson, Jim Greer, Carl Gutwin

Laboratory for Advanced Research in Intelligent Educational Systems (ARIES), Department of Computer Science, University of Saskatoon, 110 Science Place, Saskatoon, SK, Canada

ARTICLE INFO

Article history:
Received 29 July 2013
Received in revised form
4 March 2014
Accepted 6 March 2014
Available online 15 March 2014

Keywords:
Lecture recording
Lecture capture
Podcasting
Media in education
Learning analytics

ABSTRACT

The literature is mixed as to whether the addition of lecture capture technologies provide for better student success. In this work, we consider not just the broad effect of lecture capture technology on academic achievement between cohorts, but whether this effect is related to patterns of viewership among learners. At the centre of our interest is determining whether there are strategies learners take in their reviewing of content week-to-week that may result in better achievement. To investigate this, we describe a method for modelling learners based on their interactions with lecture capture systems. Unlike investigations done by others, our models emerge from the activities of the learners themselves, and are based on the results of applying unsupervised machine learning (clustering) techniques to student viewership data. These models describe five different classifications of learner interactions, and we show that one of these is positively correlated with academic achievement. We further validate our results through repeated experimentation, and describe how such models might be used by early-alert systems.

1. Introduction

Over the last ten years there has been a shift in classroom recording away from expensive, studio-based televised solutions primarily used for distance education, to inexpensive automated solutions used to augment the traditional on-campus student experience. This shift has come both because of increased expectation of video resources from students (e.g. who are used to social video technologies such as YouTube) as well as the increased ability to scale recording technologies throughout the institution (through automation technologies and cost reduction of video recording devices). Instead of small courses being offered with professional television crews and videographers supporting them, dozens of courses in different disciplines and programs can be simultaneously recorded, processed, and made available to learners using inexpensive automated systems.

This change in the use of video technology brings with it both challenges and opportunities. One of the challenges that occurs with low cost solutions is that pedagogical approach and effectiveness of the technology is often poorly understood. Since the costs are low there is minimal motivation to determine how well the technology helps students. Many deployments in the field are led by technicians as opposed to instructional designers, who are often more involved in higher-cost educational support activities. At the same time, lecture capture learning environments, like learning content management systems, bring with them unprecedented opportunities to gather information on learner behaviours. Being able to log student actions in the learning environment allows educational researchers to compare cohorts of learners to one another based on learning activity and educational outcome. Such activity has immediate benefits to instructors and students who use these technologies, as it provides a deeper understanding of whether a technology is valuable to the teaching and learning process.

While there are a number of interesting educational issues with respect to lecture capture technology, such as student motivation, classroom attendance, and pedagogical techniques, this work focuses on the issue of whether there are patterns of viewership of

* Corresponding author.
E-mail addresses: cab938@mail.usask.ca, brooksch@umich.edu (C. Brooks), gke703@mail.usask.ca (G. Erickson), greer@cs.usask.ca (J. Greer), gutwin@cs.usask.ca (C. Gutwin).

http://dx.doi.org/10.1016/j.compedu.2014.03.002
0360-1315/

recorded lectures and, if so, how these patterns correlate with academic performance. The end goal of this line of research is to determine both good and bad uses of lecture capture technology, in order to encourage students to use it appropriately. The key data we examine in this work is usage data. Created as a side effect of student interaction with learning systems, usage data describes the actions and choices students have made. We look at this data through the lens of machine learning, a computational and statistical process, which can be effective at summarizing and revealing patterns in large or complex datasets. Specifically, through the application of unsupervised machine learning, we demonstrate that there are at least five distinct patterns of lecture capture use, and that one of these (the consistent weekly viewing of lecture video) is positively correlated with higher academic performance. We go further and replicate these results on two related cohorts, showing that the effect persists despite no significant difference between the incoming grades of learners in different clusters. We go on to demonstrate how this technique can be used to form a descriptive abstract usage model. This model provides insights into how learners use lecture capture, and can be used by educational researchers and instructional designers for scaffolding of the learning process. We finally address the implications such a model may have for predicting student success for early-alert student support systems.

1.1. Related work

There has been a variety of conflicting evidence with respect to whether the use of lecture capture technologies affects academic performance. In a broad survey of the area, Heilesen (2010) cites a number studies that show no effect on performance (Abt & Barry, 2007; Baker, Harrison, Thornton, & Yates, 2008; Hodges, Stackpole-Hodges, & Cox, 2008) as well as a number that show a positive effect (up to 9.5%) depending on the kind of evaluation being given (Carle, Jaffee, & Miller, 2009; Kurtz, Fenwick, & Ellsworth, 2007; McCombs & Liu, 2007; McKinney, Dyck, & Luber, 2009; Smith & Fidge, 2008). Heilesen ends his survey considering the lack of consensus of the efficacy of lecture capture (which he refers to as podcasting) with an insightful point:

> "As matters stand, the answer to whether or not engaging in podcasting is worthwhile for purely academic reasons is not entirely clearcut. Evidence that students score better at exams after having listened to podcasts is inconclusive, and most likely the positive effects claimed should be attributed to the uses made of the technology rather than the technology per se." (pg. 1066)

It is the way lecture capture technology is used that Heilesen attributes to the change in student achievement, and this usage is the data we are interested in better understanding through the modelling of learners. Leadbeater, Shuttleworth, Couperthwaite, and Nightingale (2013) have similar goals, and clustered learners into one of two groups depending on the frequency with which they watched lectures. These groups, denoted as *high usage* and *low usage*, did not differ with respect to academic performance; however other demographic information (dyslexia and English as a second language status) did differentiate these two groups. Our study on learner usage of lecture capture technology is similar in goal but our methodology allows us to ask slightly more nuanced questions, and thus reach different conclusions. We revisit the differences between our work and Leadbeater et al. (2013) in more detail in §4.

Finally, it is useful to note that there are many studies on the use of lecture capture that approach the issue using qualitative methods. Studies that use questionnaires to elicit self-reports of increased learner satisfaction, increased flexibility of learning, and positive performance are numerous within the literature (see Pursel & Fang, 2012; for an overview of some of the approaches). We do not explicitly consider these issues here, but note that there may be interesting correlations with the way lecture capture tools are used (e.g., the clusters we identify) and the responses that learners give to various survey instruments.

2. Methodology

The lecture capture environment that we used was a predecessor of the freely available Opencast Matterhorn system (Brooks, McKenzie, et al., 2011) and, like Matterhorn, it provides the ability to track interactions learners have with the system. As learners interact with the user interface through pressing buttons, selecting list items, hovering the mouse over images, and seeking on the timeline, logs of this data are sent back to a central server. Most importantly for this work is that *heartbeats* are also logged – automatic events every 30 s that indicate the video the learner has loaded, the position the learner is at in that video, and whether that video is playing or not. From this, a rough estimate of the amount of video that has been watched can be calculated.[1]

Our experiments involved three cohorts of second year (sophomore) students in science courses at a research-intensive university. Each cohort was taught through traditional face-to-face delivery (i.e., not distance education) on a university campus. Students could attend three fifty minute lectures per week for thirteen weeks, all of which were recorded and made available online 24–48 h of the live lecture. Students also had access to various other study aids such as textbooks, online resources, laboratories, and tutorials, and each class included a midterm and a final examination, and students were notified via e-mail when lectures were posted online. The cohorts used were:

1. *Chemistry 2010 Spring*: This course was an introduction to organic chemistry and was taught in three parallel sections by different faculty members, but used a single curriculum and a shared set of assessments (common examinations and assignments). A total of 636 learners were given access to the lecture capture environment over the spring 2010 term.
2. *Chemistry 2011 Spring*: This course was an introduction to organic chemistry and was taught similarly to the 2010 offering, with a slightly modified curriculum. Multiple sections of the course were taught and used a shared set of assignments and examinations. A total of 546 learners were given access to the lecture capture environment over the spring 2011 term.

[1] There are multiple methods for creating this measure of "minutes of video watched", all of which are susceptible to different forms of error. For this work, heartbeats were considered to be indicative of 30 s of video watched if the heartbeat indicated the player was in a playing state, and all other events were discarded. Thus the total amount of time a learner spent watching video in some time period was the number of heartbeats in that time period times 30 s.

3. *Biomolecules 2011 Spring*: This course was a bio-molecules and biochemistry course which has similar prerequisites to the Chemistry courses. There was one section of this course taught, and 197 learners were given access to the lecture capture environment over the spring 2011 term.

Participation in this study was voluntary and no remuneration was given. If learners used the lecture capture system, their usage data was collected in a pseudonymous form. Usage data was summarized into weekly viewing habits: Learners who watched at least five minutes of video lecture content in a calendar week were considered to have watched content in that week, whereas learners who watched less than five minutes of content that week were considered to have not watched any content that week. This time limit was used to reduce the novelty-effect bias associated with having a new tool in the learning environment.

The main technique used in data analysis was unsupervised machine learning. Unsupervised machine learning represents a class of statistical methods designed to quantify the structure of a set of data based on the attributes of the data alone. In the experiments described in this paper, we used the *k*-means clustering method to partition the data set into *k* different groups. In *k*-means, *k* random data points are chosen by the algorithm as cluster centroids, and all other instances are assigned to the cluster which minimizes the Euclidean distance between the centroid of that cluster and the given instance. The centroid for a cluster is then recalculated as the means of the instances assigned to that cluster. This process repeats until cluster centroids do not change. Due to the sensitivity of *k*-means clustering to local minima, we followed common practices and repeated this process of choosing *k* new random starting points 1000 times, and chose the best fitting centroids.[2]

Choosing an appropriate value for *k* is a difficult problem, and one that is under investigation in the data mining community. Choosing a very large number for *k* (e.g. one that approaches the number of instances being classified) increases how specific clusters are, and reduces the generalizability of the model since many (or every) instance is in its own cluster. Choosing a very small number for *k* increases the amount if disagreement (which we refer to as error later in this paper) between instances and the model, reducing the accuracy of a given cluster. For all experiments described in this paper we have assigned $k = 5$ as is described more fully in our previous work (Brooks, Epp, Logan, & Greer, 2011).

In this study, we were interested in understanding whether we can (1) model the viewing habits learners have with lecture capture technology and (2) correlate those habits to academic achievement data. We are interesting in using such models to both describe undergraduate learning activity, and to predict student success based on their activity. To do this, we first applied unsupervised machine learning to the Chemistry 2010 dataset (described in §3.1) to build a model of student activity. From this, we generalized to an abstract model that allowed us to classify learners in other similar future courses with different examination dates. We validated this abstract model on the Chemistry 2011 dataset (§3.2). This abstract model was then applied to a similar domain (Biomolecules), where we quantified the levels of error the model creates (§3.4). This error calculation, described in more detail further on, allows us to quantify how similar a given group of learners are to the abstract model. Finally, we discuss the issues in using this approach for the predictive modelling of more generalized learner achievement based on lecture capture usage (§3.5).

3. Results

3.1. Building the usage model

The weekly viewing habits of learners who used the lecture capture tool ($n = 232$ of the 636 students who had access) from the *Chemistry 2010* cohort were used as attributes with *k*-means clustering ($k = 5$) to create a model of learner behaviours. A limit of five clusters was chosen based on expert estimates of how learners might use the system. It is important to note that the *k*-means algorithm does not use the content of the hypotheses themselves when generating clusters, and instead generates clusters based on statistical similarity of the participant data. The attributes used to form clusters where learners watched (w) or did not watch (d) at least five minutes of lecture video in a given week.

The output of the clustering process is shown in Table 1. Each learner is fitted into the cluster which minimizes the differences between the participant behaviour and the cluster centroids. Each row in the table indicates a cluster as determined by the *k*-means process, and the centroid for that cluster are the values of watching (w) or not watching (d) content in a particular week. There were 16 weeks in the course, but week seven was omitted from the modelling activity as it was a university-wide midterm break. Weeks 1 through 13 were those that had regular lecture activity, and weeks 14 through 16 denote weeks between the end of lectures and the final exam (e.g., examination study period). Clusters range in size from 8 to 103 participants, with mean error between 5% and 25%.

Each cluster was given a descriptive label by ourselves based on the pattern we observed. The first group in this model has learners who habitually watch lectures throughout the term, so they were labelled *High Activity* learners. The second group is made up of learners who observed the lecture the week before the midterm examination, so they were labelled as *Just-In-Time* learners. The third and fourth groups appear to correspond (roughly) to consistent viewership in either the first or second half of the course, so they were labelled as *Early* and *Deferred* learners respectively. Finally, the last group is made up of learners who habitually did not watch lectures. To be included in the study, learners must have watched at least five minutes of video in some week, but the overall pattern of behaviour from these *Minimal Activity* learners suggests that the tool is used sparingly throughout the course instead of primary study aid, perhaps to catch up on missed lectures. We excluded 404 students from the study because they did not watch more than 5 min of video in at least one week. The time between the end of lectures and the final exam revealed no discernible pattern and did not seem indicative of the clusters as a whole, and was discarded from the general model.

[2] There are various methods of determining an appropriate number of iterations to run and, given the size of the data set, it was reasonable to choose a large set of iterations to minimize issues of local minima. It is important to note that the problem *k*-means attempts to solve is an NP hard problem (Dasgupta, 2008), thus an exhaustive search of the clustering space is not feasible. For this work we used a random starting seed with *k*-means clustering and allowed the algorithm to complete 500,000 iterations before stopping. We repeated this process 1000 times, and report here the best fitting result.

Table 1
K-means clustering of *Chemistry 2010* lecture capture data ($k = 5$). Clusters range in size from 8 to 103 participants, with mean error between 5% and 25%. A *w* indicates that students in this group tended to watch videos while a *d* indicates that students in this group did not tend to watch videos in a given week.

		Academic week														
		Regular lecture period												Exam period		
		1	2	3	4	5	6	8	9	10	11	12	13	14	15	16
High Activity	Centroids	*w*	*w*	*w*	*w*	*w*	*w*	*w*	*w*	*w*	*w*	*w*	*w*	*w*	*d*	*d*
$n = 10$, $\bar{x}_{error} = .25$	Error	.50	.30	.20	.20	.10	.30	.00	.10	.30	.20	.20	.40	.30	.20	.40
Just-In-Time	Centroids	*d*	*d*	*d*	*d*	*d*	*d*	*d*	*w*	*d*	*d*	*d*	*d*	*d*	*d*	*d*
$n = 98$, $\bar{x}_{error} = .05$	Error	.03	.08	.08	.04	.05	.01	.10	.00	.02	.05	.04	.04	.05	.05	.15
Early	Centroids	*w*	*w*	*w*	*w*	*w*	*d*	*d*	*d*	*d*	*d*	*d*	*d*	*d*	*d*	*d*
$n = 8$, $\bar{x}_{error} = .15$	Error	.38	.00	.25	.12	.38	.12	.00	.25	.00	.12	.00	.00	.00	.25	.38
Deferred	Centroids	*d*	*d*	*d*	*d*	*d*	*d*	*w*	*w*	*d*	*w*	*w*	*w*	*w*	*d*	*d*
$n = 13$, $\bar{x}_{error} = .21$	Error	.08	.31	.15	.00	.08	.08	.38	.23	.38	.31	.08	.15	.31	.23	.31
Minimal Activity	Centroids	*d*	*d*	*d*	*d*	*d*	*d*	*d*	*d*	*d*	*d*	*d*	*d*	*d*	*d*	*d*
$n = 103$, $\bar{x}_{error} = .05$	Error	.02	.06	.04	.02	.06	.02	.08	.00	.07	.08	.05	.02	.03	.02	.19

Learners are placed into clusters based upon which cluster best represented their activity. For instance, a learner who watches video every week except for the first and sixth weeks will be placed into the *high activity* cluster. Despite this learner not fitting perfectly with this cluster (which would require that they watch at least five minutes of video every week), his or her activity patterns are most similar to the centroid that describes the cluster. Similarly, a learner who watches during the midterm week and week 12 will be placed in the *just-in-time* cluster. That the learner watched lectures during week 12 contributes to the error the cluster will have. Thus the centroids are not the only interesting aspects of the clusters, but the error is as well. Table 1 shows the amount of disagreement learner attributes in a cluster have with the centroid of that cluster (*error*). Some weeks have large disagreement (e.g., week 1 of the *high activity* cluster, where 50% of the learners in that cluster did not watch video, or weeks eight and nine of the *deferred* cluster where 38% and 23% of the learners in that cluster did not watch video), while others have perfect agreement (e.g., week eight in the *high activity* cluster).

Error for a particular cluster (denoted $\bar{x}_{error}$) refers to the mean difference between the centroid for that cluster and the set of learners who best fit that cluster (e.g., a ratio of disagreement). This is calculated by comparing each learner to the centroid of the cluster they are associated with, and calculating a difference between the learner actions and the centroid actions, giving a value in the range 0 (perfect fit) to 1 (complete disagreement). Eqn. (1) describes the method for calculating this, where n is the number of students, w is the number of weeks data exists for, s is the set of students being considered, and cluster (x) is the list of cluster centroids that the student best fits. Error values for individual weeks of usage are also given in Table 1, and the overall mean error for the model is 5.37%.

$$\frac{\sum_{x \in s}\left[\sum_{i=0}^{w}\left(|x_i - cluster(x)_i|\right)\right]}{n^{*}w} \tag{1}$$

The results of the clustering activity provide a number of insights into learner activities. For instance, some learners only watch lectures during the week of the midterm (the second row of Table 1, labelled *Early*), while others watch fairly regularly (the first row one of Table 1, labelled *High Activity*). The two largest groups by far are those who watch during the week of the midterm, and those who watch minimal video otherwise, accounting for a total of 201 of the 232 students (86.6%) who used the system. Further, while there were distinct patterns of use among learners, the error for certain weeks of the academic schedule call into question the usefulness of including those weeks in a more general model. For instance, the last two weeks of the course (corresponding to the time between the end of classes and the final examination, columns labelled 15 and 16) tended to have a high amount of disagreement between participants and the centroids. The disagreement for these weeks, ranging from 19% to 40%, suggests that activity throughout the term is not indicative of behaviours between the end of term and the final exam.

From this data it is possible to build a more abstract model of viewing habits of learners, relating cluster membership with weekly watching behaviour. The goal of developing such a model is to be able to accurately categorize learners in a course into groups with behaviours corresponding to pedagogically meaningful patterns carrying labels like those given in Table 1. Such abstract models can then be validated across multiple offerings of a course, disciplines, and year of study to understand the effect each of these clusters has on how learners use learning tools. The models can also be correlated with learning outcomes, such as final grade or perceived value of a course. Correlations may be useful in determining whether lecture capture is a useful tool for appropriate technology to deploy for a given discipline, instructor, or learner. Using the data from the *Chemistry 2010* cohort (Table 1), we developed an abstract model for five clusters (Table 2) of activity and labelled them as previously described. In doing so, our goal was for the model to fit well with the original clustered data, yet be general enough to be applied to similar courses.

Table 2
A general model for lecture capture use shown as it would apply to a thirteen week term. The two week final examination period does not factor into this model, and thus the model ends with week 13. Week 8 represents a calendar week that ends with a midterm examination.

	Academic week												
	1	2	3	4	5	6	7	8	9	10	11	12	13
High Activity	*w*	*w*	*w*	*w*	*w*	*w*	*w*	*w*	*w*	*w*	*w*	*w*	*w*
Just-In-Time	*d*	*d*	*d*	*d*	*d*	*d*	*d*	*w*	*d*	*d*	*d*	*d*	*d*
Early	*w*	*w*	*w*	*w*	*w*	*w*	*w*	*w*	*d*	*d*	*d*	*d*	*d*
Deferred	*d*	*d*	*d*	*d*	*d*	*d*	*d*	*w*	*w*	*w*	*w*	*w*	*w*
Minimal Activity	*d*	*d*	*d*	*d*	*d*	*d*	*d*	*d*	*d*	*d*	*d*	*d*	*d*

Table 3
Quantification of model error for the Chemistry 2010 and 2011 data sets. The $k = 5$ column indicates the amount of error that exists when running the *k*-means process with a limit of five clusters. The Model column indicates the amount of error that exists when using the abstract model described in Table 2. In both cases, error was calculated using Eqn. (1). The final column indicates the net gain in error when using the abstract model instead of the *k*-means clustering process.

	n	$k = 5$	Model	Net error difference
Chemistry 2010	232	.063	.067	.005
Chemistry 2011	333	.099	.109	.010

While the abstract model is inspired by the results of the unsupervised machine learning process, the two models are not identical. For instance, there is some overlap with the learners that used the lecture capture tool in the second half of the semester and those who used the tool only for preparation for the midterm examination (columns 8 and 9 of Table 1 in the *Deferred* row). We hypothesized that the learning strategy being employed by learners was that the midterm examination prompted a realization that more learning support was needed and, once the learner had tried the lecture capture system as a study aid, satisfaction was high and use of the tool continued in a regular fashion. Thus our abstract model reflects consistent use by the *Deferred* group after the midterm examination despite the one confounding week of data in Table 1 (week 10, with a high level of error). The motivations behind the third cluster, the *Early* learners, was somewhat less clear. This cluster is extremely small ($n = 8$), and watched video for the first five weeks, then didn't for the rest of the term. However, the error rises in the week of the midterm examination, indicating that two of the learners (25%) in this cluster did return to the tool to aid in their review. We set the abstract model for this group as consistent access to the tool throughout the first half of the course under the belief that students were attempting to develop consistent study habits then leaving the course after a disappointing midterm exam. We have identified this group as one that would benefit from extra data analysis (e.g. enrolment changes).

The discretization of the data into week-long time periods makes analysis sensitive to alignment issues with respect to the day of the week. There was some evidence to suggest that access to lecture capture by learners is cyclical with respect to day of week. Previous work (Brooks, Epp, et al., 2011) ignored day of week, but for comparison and generalization of models between academic terms it seems more principled to consider weekly accesses as starting on the first working day of the calendar week (as is done here). The high levels of error in these weeks for this group (23–38%) suggests that if there is an effect it is limited, and we chose to define the categories of *Early* and *Deferred* learners around the midterm exam date.

3.2. Validating the usage model

We claim that the abstract model presented is robust enough to represent learner viewership activity in courses similar to that of the *Chemistry 2010* cohort. With any model a certain amount of error may be present, and it is up to the investigator to consider whether the error is appropriate for their intended use of the model. For instance, a predictive warning system that used this form of model to send reminders to learners to watch lecture videos might be of limited risk and thus a very high amount of error (e.g., >20%) might be reasonable. If this model was to be used to determine aptitude for entrance into a professional program, a very small amount of error (e.g. <1%) might be more appropriate.

There are two sources of error that happen with this modelling technique: The limiting of $k = 5$, and the application of the general model to a course that it does not fit well. The first of these can be calculated as described in Eqn. (1), and high values suggest that a new model with a different number of centroids should be calculated. The second of these can be obtained by running the clustering methods described in the previous section to calculate the error in the model formed when $k = 5$. Then the error for the abstract model (Table 2) can be calculated, and the difference between these values can be attributed to the application of the general model. This value of error describes how well the general model fits the cohort and whether another structure may be more appropriate. In this second case, the value of *k* must be held constant in order for the results to be valid.

Table 3 shows values of the general model using the original cohort upon which the model was formed (*Chemistry 2010*), and a cohort from the following year (*Chemistry 2011*). The overall average disagreement went up between the two offerings of the course, suggesting that additional investigation using different values of *k* may be warranted. However, this increase in error was quite small, only 3.6%. More interestingly is the net error difference when applying the abstract model to the data instead of *k*-means. In both cases, the error difference is extremely small (0.5% for the *Chemistry 2010* cohort and 1.0% for the Chemistry 2011 cohort). As the abstract model was based on the clustering of the *Chemistry 2010* data we would expect there to be minimal error. That the error is low for the following offering of the course suggests that the abstract model is a robust instrument for describing student interactions across offerings of this Chemistry course.

3.3. Correlation with grades

Instructional goals are often represented by midterm and final examinations test items and, while grades are not a complete measure of learning, they are often used by learners and instructors as a proxy for learning. Correlating patterns of behaviours with differences in grades

Table 4
Midterm, final examination, and overall grade averages and standard deviations broken down by cluster in percentages for the Chemistry 2011 course. Students who did not use the system but received a midterm and final examination grade are described in the last category, *Non-Users*.

Cluster label	*n*	$\text{Midterm}_{\bar{x}}$	Midterm_{σ}	$\text{Final}_{\bar{x}}$	Final_{σ}	$\text{Overall}_{\bar{x}}$	Overall_{σ}
High Activity	14	77.32	15.71	75.43	22.19	80.14	14.49
Early	18	64.71	15.72	58.98	17.81	66.82	13.85
Just-In-Time	86	68.14	15.92	60.49	23.68	70.69	14.39
Minimal Activity	191	64.30	15.36	58.98	22.89	68.41	14.71
Deferred	24	63.33	12.20	59.83	20.21	69.04	12.43
Non-Users	196	65.75	15.59	62.92	19.75	69.42	14.23

Table 5
Dunnett–Tukey–Kramer confidence values between pairs of clusters using midterm examination marks for the *Chemistry 2011* cohort. The means differences between groups is small for all pairs except for those that include the *High Activity* learners, which have a substantial differences in examination mark in the 9.18%–13.99% range. Despite this, no statistical significance is found at the $p = 0.05$ level, likely due to the small size of the *High Activity* cluster.

	High Activity		Early		Just-In-Time		Minimal Activity		Deferred Activity	
	95% CI	$\Delta_{midterm}$	95% CI	$\Delta_{midterm}$	95% CI	$\Delta_{midterm}$	95% CI	$\Delta_{midterm}$	95% CI	$\Delta_{midterm}$
Early	−31.8, 6.6	12.61%								
Just-In-Time	−24.6, 6.2	9.18%	−10.6, 17.5	3.43%						
Minimal Activity	−27.2, 2.3	13.02%	−13.2, 13.5	0.17%	−10.2, 3.7	3.26%				
Deferred	−30.5, 2.5	13.99%	−16.6, 13.9	1.37%	−15.0, 5.4	4.80%	−10.7, 7.6	1.54%		
Non-Users	−26.3, 3.1	11.67%	−12.2, 14.3	1.05%	−9.2, 4.4	2.38%	−4.3, 6.0	0.88%	−11.4, 6.6	2.42%

Table 6
Dunnett–Tukey–Kramer confidence values between pairs of clusters using final examination marks for the *Chemistry 2011* cohort. Similar to midterm examination marks, the means differences between groups is small for all pairs except for those that include the *High Activity* learners, which have a substantial differences in examination mark in the 12.51%–15.60% range. Despite this, no statistical significance is found at the $p = 0.05$ level.

	High Activity		Early		Just-In-Time		Minimal Activity		Deferred Activity	
	95% CI	Δ_{final}	95% CI	Δ_{final}	95% CI	Δ_{final}	95% CI	Δ_{final}	95% CI	Δ_{final}
Early	−38.4, 5.5	16.45%								
Just-In-Time	−31.8, 6.2	12.75%	−10.8, 18.1	3.69%						
Minimal Activity	−32.1, 4.5	13.81%	−10.9, 16.1	2.63%	−9.0, 6.9	1.05%				
Deferred	−37.1, 6.0	15.60%	−16.8, 18.5	0.85%	−16.8, 11.1	2.84%	−14.7, 11.1	1.78%		
Non-Users	−20.7, 5.7	12.51%	−9.5, 17.3	3.94%	−7.6, 0.8	0.24%	−4.5, 7.1	1.30%	−15.8, 9.6	3.08%

can provide evidence of detecting learning from activity. In our studies, learners use lecture capture as one tool to aid in learning, but many other tools and methods contribute to learning (e.g., online quizzes, in class lectures, textbooks, study groups), which makes identifying the effect of any single tool difficult. That is, we would expect it to be very difficult to obtain statistically significant results given that learners were only intrinsically motivated and used a variety of tools to aid in their learning. To aid in the understanding of our results, we have included statistics values as confidence intervals in Tables 5–8. While no values are statistically significant at the $p = 0.05$ level, most of the comparisons of the *High Activity* to other groups are heavily skewed towards the negative, suggesting that there may be an underlying weak effect present. All analyses were done with the Dunnett–Tukey–Kramer test, a pairwise test for significance for unbalanced datasets.

The effect size of being categorized in a given cluster can be quite large. An analysis of average marks and standard deviations (shown in Table 4) shows that learners categorized in the *High Activity* cluster outperform their peers by up to 16.45%, while learners in other clusters have more or less same achievement as one another. Pairwise Dunnett–Tukey–Kramer tests and means differences for the midterm examination (Table 5), final examination (Table 6), and overall course grade (Table 7) for the *Chemistry 2011* cohort were calculated to contextualize this further. These tables show both the range of the level of statistical significance (the confidence interval, CI) as well as the difference of the marks between groups (Δ). For instance, the first column labelled $\Delta_{midterm}$ in Table 5 indicates the difference in midterm examination mark between students in the *High Activity* cluster and those in other clusters. Across all three evaluation scores (midterm, final, and overall marks), there are large differences between those learners in the *High Activity* cluster and other cluster (ranging from 9.18% to 16.45%), and minor differences between any other two groups (from 0% to 4.80%). While these tests do not show statistical significance at the $p = .05$ (that is, the range of the confidence interval for each comparison includes the null hypothesis, that there is 0 difference) there is a general trend of negatively skewed confidence intervals for comparisons including the *High Activity* cluster. It thus appears that there may be some evidence to suggest that the marks for the *High Activity* cluster of learners are different than those of the other clusters, and more investigation may be warranted.

The effect between being in the *High Activity* cluster and marks is a positive one, and being placed in this cluster indicates likelihood of a higher mark. The question that arises is whether this is a causal relationship or only a correlative one. That is, does regular viewing of lecture capture videos help learners achieve higher grades, or are learners who get higher marks just more likely to view lecture capture video regularly. Answering this question is difficult given the lack of controls in our study – learners in the *High Activity* cluster may have other traits that differentiate them from the students in other clusters including gender, ethnic background, workload, or academic experience. However, we did investigate incoming grade point average of the learners in the study to investigate whether previous grades had an impact (e.g. that historically high-achieving students self-select to use lecture capture (Table 8)). Not only is there no statistically significant

Table 7
Dunnett–Tukey–Kramer confidence values between pairs of clusters using overall course mark for the *Chemistry 2011* cohort. The means differences between groups is small for all pairs except for those that include the *High Activity* learners, which have a substantial differences in examination mark in the 9.18%–13.28% range. Despite this, no statistical significance is found at the $p = 0.05$ level.

	High Activity		Early		Just-In-Time		Minimal Activity		Deferred Activity	
	95% CI	$\Delta_{overall}$	95% CI	$\Delta_{overall}$	95% CI	$\Delta_{overall}$	95% CI	$\Delta_{overall}$	95% CI	$\Delta_{overall}$
Early	−38.9, 2.2	13.28%								
Just-In-Time	−21.8, 2.5	9.61%	−8.1, 15.4	3.67%						
Minimal Activity	−22.5, 0.6	11.00%	−8.8, 13.5	2.32%	−7.1, 4.4	1.24%				
Deferred	−23.9, 2.8	10.57%	−10.2, 15.7	2.7%	−9.7, 7.8	0.96%	−7.6, 8.4	0.38%		
Non-Users	−21.5, 1.5	10.03%	−7.8, 14.3	3.25%	−6.0, 5.2	0.42%	−3.3, 5.1	0.92%	−8.3, 7.2	0.54%

Table 8
Dunnett–Tukey–Kramer confidence values between pairs of clusters using incoming grade point averages for the *Chemistry 2011* cohort. Only 225 learners are considered (142 who used the lecture capture tools and 83 non-users) as incoming grade point average is not available for all learners. Most interesting is the first column which compares the *High Activity* learners to other groups of learners, showing no differences of statistical significance and in most cases small means differences in incoming mark.

	High Activity		Early		Just-In-Time		Minimal Activity		Deferred Activity	
	95% CI	Δ_{in}	95% CI	Δ_{in}	95% CI	Δ_{in}	95% CI	Δ_{in}	95% CI	Δ_{in}
Early	−38.6, 25.5	6.55%								
Just-In-Time	−26.5, 27.3	0.40%	−12.0, 25.9	7.00%						
Minimal Activity	−26.8, 25.9	0.60%	−12.1, 24.0	5.94%	−10.1, 8.1	1.01%				
Deferred	−41.0, 24.0	8.54%	−27.0, 23.0	2.00%	−28.6, 10.7	8.95%	−26.3, 10.4	10.40%		
Non-Users	−26.0, 26.3	0.17%	−11.3, 24.8	6.72%	−9.3, 8.9	0.24%	−6.2, 7.7	0.77%	−10.1, 27.6	8.71%

evidence to support this hypothesis, but the confidence intervals comparing the *High Activity* students to others are generally well balanced (unlike the skewed CI comparing midterm, final, and overall marks), and the means differences are minimal in three of the five comparisons ($\leq$0.60).

3.4. Applying the usage model to related domains

Forming a model on the *Chemistry 2010* dataset and verifying its utility on the *Chemistry 2011* dataset provides a baseline that might be adapted to other courses. To show cross-course validity of this model, data was collected from a second year Biomolecules course in 2011. This course has many of the same attributes as the *Chemistry 2010* and *Chemistry 2011* courses: it was made up of a large cohort of learners ($n = 190$), had a single midterm, and required some of the same prerequisites. The course was taught by a single instructor, and all students were in a single section.

The first step in understanding the suitability of the general model to this course is to calculate the amount of error and compare it to the previous results (Table 9). The *Biomolecules 2011* data set does not cluster as well to five clusters compared to the *Chemistry 2010* or *Chemistry 2011* values. Whereas the Chemistry courses had $k = 5$ values with under .1 error (.063 and .099 respectively), the Biomolecules $k = 5$ cluster had .133 error, which corresponds to an average of roughly one misclassified week of interaction per learner. Despite this, the application of the five high level cluster descriptions in the model to the Biomolecules courses introduced minimal additional error (.0298), suggesting that if keeping with $k = 5$, the general model is a reasonable fit.

While the Biomolecules 2011 course had midterm and final exams similar to the Chemistry 2011 course, only the overall grade of learners was available for analysis. This grade includes the aggregate of examinations, assignments, and laboratory exercises, and may be curved or scaled unbeknown to ourselves. Nonetheless, it is worthwhile considering Tukey HSD values to compare clusters to see how well the model performs (Tables 10 and 11). The pattern of p-values (e.g., lower for the cluster of *high activity* versus all other clusters) is similar to that of the Chemistry 2011 final marks (Table 7), though there is no statistically significant difference between groups at the $p = .1$ or $p = .05$ levels. Further, it is interesting to note that the pattern of means differences are also similar to those of the *Chemistry 2011* cohort – learners who watch lecture video regularly (the *High Activity* cluster) consistently achieve 10% or higher on formal evaluations.

3.5. Using the model to predict achievement

Beyond identifying the impact lecture capture technologies have on achievement, it is possible to use the interaction data of learners to attempt to predict educational outcomes while a course is being offered. Early warning of academic problems allows for academic interventions, either by experts (e.g., referring a learner to a tutor) or automated (e.g., the motivational power of the *Course Signals* system (University of Purdue, 2012)). To build a predictive model on lecture capture interaction data it is important to know what categorizational power exists at any particular time. For instance, is it appropriate to use this model only at the end of the course for reflecting on learning, or can it be used by an instructor during the teaching period? Is it useful to use this model in the first week of a course, or is it only useful after

Table 9
Comparison of mean disagreement between an ideal clustering and a pedagogical model within the Biomolecules 2011 data set. The values can range between 0 (perfect fit) and 1 (complete disagreement), and are given by the equation described in Eqn. (1). Ideal clusters were restricted by $k = 5$.

	n	$k = 5$	Model	Net error difference
Biomolecules 2011	180	.133	.163	.0298

Table 10
Overall grade averages and standard deviations broken down by cluster in percentages for the *Biomolecules 2011* course. Grades for non-participants of the lecture capture system were not available for comparison.

Cluster label	n	Overall$_{\bar{x}}$	Overall$_{\sigma}$
High Activity learners	15	78.53	12.20
Early learners	10	66.5	15.67
Just-In-Time learners	48	68.58	17.82
Minimal Activity learners	92	67.47	18.82
Deferred learners	15	67.6	16.05

Table 11
Tukey HSD confidence values between pairs of clusters using overall course marks for Biomolecules 2011. While no statistical significance is found at the $p = .10$ or $p = .05$ levels, the general trend of low p values for the *High Activity* cluster is similar to that found in *Chemistry 2011* dataset (Table 7).

	High Activity		Early		Just-In-Time		Minimal Activity	
	p	$\Delta_{overall}$	p	$\Delta_{overall}$	p	$\Delta_{overall}$	p	$\Delta_{overall}$
Early	.472	12.03%						
Just-In-Time	.335	9.95%	.997	2.08%				
Minimal Activity	.179	11.06%	.9998	0.97%	.997	1.11%		
Deferred	.456	10.93%	1.000	1.10%	.9997	0.98%	1.000	0.13%

midterm examinations? This section of the work will investigate how prediction accuracy of the model changes as more interaction data about learners is collected.

Table 12 shows the prediction accuracy of cluster membership using the *Chemistry 2011* cohort. At each week five values are given for each cluster; a) the number of learners who are predicted to be in a given cluster by the end of the term and who, from their activity, already best fit this cluster (*true positives*), b) the number of learners who are predicted to be in another cluster by the end of the term and whose data thus far suggests they do not fit this cluster well (*true negatives*), c) the number of learners who are predicted to be in this cluster by the end of the term but whose data thus far suggests they best fit another cluster (*false negatives*), d) the number of learners who are predicted to be put into this cluster by the end of the term but whose data suggests they do not fit this cluster the best thus far (*false positives*), and e) the number of learners who are predicted as best fitting into this cluster (either as a *true positive* or a *false positive*) but who could optimally fit into at least one other cluster (*borderline*).

The clustering model developed uses a series of attributes relating to the week of lecture with a binary value of either watched or not-watched. It is impossible to differentiate clusters after the first week of data alone, as several clusters have the same pattern of usage (e.g., the pattern for the *High Activity* and *Early* learners is denoted by watching the first week of lecture, while the activity pattern for the *Just-In-Time*, *Deferred*, and *Minimal Activity* learners are all denoted by not watching the first week of lecture). At the seventh week, it is possible to differentiate the *just-in-time* learners from the *minimal activity* learners, and by week eight it is possible to differentiate all of the clusters from one another.

We introduce the term "actionable" to denote when it would be reasonable for an early alert system to intervene and recommend a learner change their behaviour. If there is sufficient confidence that a learner falls into one of the identified clusters as the semester progresses, then it is reasonable to take action and give advice about the possible consequences a learner faces and recommended behavioural changes. Depending on the instructional intervention being instigated, different values of how actionable the data is (Table 12) may be interest. For instance, if it is in week five and the instructor is planning to e-mail learners who do not fit into the *high activity* cluster to

Table 12
Actionability table for Chemistry 2011 data. At each week of instruction a learner could have either watched online video or not, and is classified into the appropriate cluster based on the model derived from Table 1. For each week of instruction, all learners can be classified as to which cluster they best fit based on their data thus far. For example, in week one a total of 70 learners best fit the *High Activity* cluster (*true positive* + *false positive*), and 11 of these learners will end up in this cluster at the end of the term based on their interactions throughout the term (*true positive*). Fifty-nine of these learners will end up in another cluster by the end of the term (*false positive*). Further, a total of 262 learners do not fit this cluster based on their interactions thus far (*true negative* + *false negative*), with the vast majority (259) not fitting this cluster by the end of the term (*true negative*). Only three learners (*false negative* do not fit this cluster after one week of data, but will by the end of term. Since it is impossible to differentiate at week one between the *High Activity* and *Early* clusters, they both have identical values for *borderline*.

		Academic week											
		1	2	3	4	5	6	7	8	9	10	11	12
High Activity	True positive	11	12	10	14	13	14	14	8	12	12	14	14
	True negative	259	238	292	282	297	293	300	314	309	313	311	318
	False positive	59	80	26	36	21	25	18	4	9	5	7	0
	False negative	3	2	4	0	1	0	0	6	2	2	0	0
	Borderline	70	92	36	50	34	39	32	4	14	4	11	3
Deferred	True positive	22	26	25	27	25	27	15	8	14	18	21	27
	True negative	65	20	34	26	32	22	213	289	288	294	297	305
	False positive	240	285	271	279	273	283	92	16	17	11	8	0
	False negative	5	1	2	0	2	0	12	19	13	9	6	0
	Borderline	262	311	296	306	298	310	107	15	23	14	15	14
Minimal Activity	True positive	164	195	191	197	196	202	202	202	199	202	197	202
	True negative	32	14	25	21	28	22	129	129	129	129	130	130
	False positive	98	116	105	109	102	108	1	1	1	1	0	0
	False negative	38	7	11	5	6	0	0	0	3	0	5	0
	Borderline	262	311	296	306	298	310	0	12	0	10	0	11
Early	True positive	11	15	13	16	15	18	18	15	18	13	18	18
	True negative	255	237	291	280	295	293	300	308	308	312	311	314
	False positive	59	77	23	34	19	21	14	6	6	2	3	0
	False negative	7	3	5	2	3	0	0	3	0	5	0	0
	Borderline	70	92	36	50	34	39	32	7	17	6	14	7
Just-In-Time	True positive	82	91	88	90	88	92	92	86	92	85	92	92
	True negative	60	20	32	24	30	22	225	231	231	237	234	240
	False positive	180	220	208	216	210	218	15	9	9	3	6	0
	False negative	10	1	4	2	4	0	0	6	0	7	0	0
	Borderline	262	311	296	306	298	310	107	7	26	6	18	7

encourage them to watch more video, then the instructor should be aware that a portion of the learners they are not emailing (in this case, twenty-one, which is the value for week five for the *high activity false positives*) will actually end up in the *early* cluster. Similarly, if it is week nine and the instructor has asked tutorial assistants to get in contact will all of the *early* learners, they should be aware that several of these learners (in this case six, given by the value for the *early false positives*) will not best fit this cluster by the end of the term. Thus it is important to contextualize that, with this model as a predictor, there is error that depends in part on the purpose for which an intervention is being made.

Although this error cannot be measured accurately until the end of the term, there are some broad comments that can be made. The *false* values drop significantly after clusters can be disambiguated from one another (Fig. 1). As mentioned, in week seven it is possible to differentiate the *just-in-time learners* from the *minimal activity learners*, and the total *false positives* drop dramatically. In week eight it is possible to differentiate between the *just-in-time* and *deferred* learners, which reduces the *borderline* value for the *just-in-time* cluster. These are two different kinds of error; in the first case, the error is a mis-prediction caused by inaccuracy in the model. The model does not fit a particular learner well, and the solution may be to include more attributes in the model or form a larger set of models (increase *k*). In the second case, the error is caused by cluster ambiguity; the clusters are too similar to one another to distinguish where a particular learner goes. The solution for this problem is the same as the first; ambiguity is not necessarily a problem depending on the instructional intervention being taken. For instance, in the first example given previously where an instructor wanted to e-mail all learners who are not in the *high activity* cluster, it does not matter if he or she is unable to distinguish between the *minimal activity* cluster or the *just-in-time* cluster as both clusters contain learners of interest.

4. Discussion

4.1. Relationship to other work

In particular, the work of Leadbeater et al. (Leadbeater et al., 2013) shares a similar goal to our own. In their work, they divided students into two cohorts based on how frequently students used the lecture capture system, and labelled these cohorts as *high usage* and *low usage*. They found no statistically significant difference between the cohorts with respect to academic performance, though they noted that the *high usage* groups tended to have a higher incidence of dyslexia and as English as a second language learners. The question that arises is why

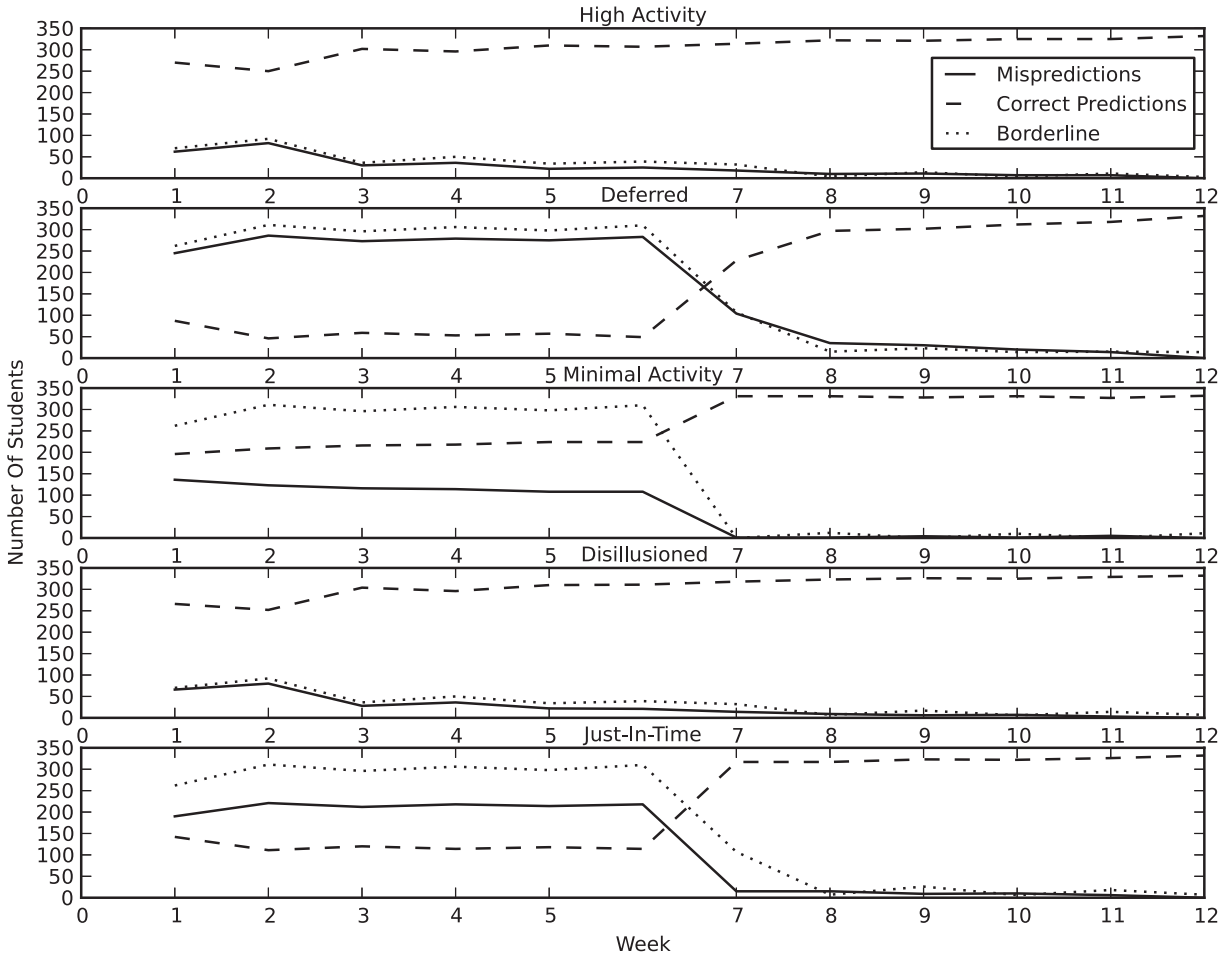

Fig. 1. Graphs plotting the prediction accuracy of learners versus week in the term. Generally, as more information is collected about learners, the correct predictions increase and the mispredictions decrease. Borderline learners are those who fit equally well in multiple clusters based on their activity.

our group of high activity learners were observed as having significant differences in academic performance while the high usage learners from Leadbeater et al. did not. The answer comes from various differences between our experiments, which we have identified as the following:

1. High usage learners are not the same as high activity learners. In our experiments, we used whether learners watched lecture video for five minutes or more in a given week as a cut-off to determine participation. A high activity learner is one who fits best with the *High Activity* cluster, which does not require watching of video every week. In Leadbeater et al., a high usage learner is one who used the lecture capture system for at least five hours over the academic semester. It is possible in our experiment for a learner to be classified as high activity even if they have watched only 35 min of video, as long as they watched this video in 5 min intervals in different weeks throughout the semester. Similarly, a learner who watched five hours of video would be classified as a *Minimal Activity* if, for instance, they watched this video all in the first week of the academic semester. In short, we use different measurements for high/low activity.
2. Frequency of use of lecture capture by instructors. In our experiments the instructors used lecture capture regularly for every lecture of the course. In Leadbeater et al. instructors could pick and choose which lectures they would record, and many modules were not recorded. For learners who form study habits early on (as our *High Activity* learners did), this difference may have caused a change in which tools a learner uses for support.
3. Small sample sizes. In our experiments we used courses with large numbers of students ($180 \leq n \leq 333$) which is particularly important given the small size of some of our clusters. Leadbeater et al. used smaller courses ($n = 76$), which may have hampered the ability to gain statistical significance given that both sets of studies were field studies and learners had access to various tools to aid in their studying. In both cases, replication studies were used to confirm results.
4. Self-reporting data. While both sets of studies used actual performance outcomes (versus self-reported outcomes), our studies included all learners registered in the course, while Leadbeater et al. asked individual student permission for assessment data. This reduced the set of participants in Leadbeater et al. to roughly half, and may have led to a self-selection bias among students. Further, Leadbeater et al. ask students to report on their use of lecture capture tools, instead of directly measuring it as we do.
5. Assessment differences. Both sets of studies looked at final examination grades, however, the cohort discussed in Leadbeater et al. did not have a midterm examination, which we see impacting student use of lecture capture (the *just-in-time* group of students).
6. Domain or discipline differences. While our studies looked at learners in Chemistry and Biomolecules courses, Leadbeater et al. observed Medical students. Anecdotally, many of the learners in the Chemistry and Biomolecules courses are pre-medicine students, so we anticipate that if this difference led to an effect it is minimal.

Perhaps most importantly, our work is different from others in that the groups of learners discussed are formed statistically through machine learning based upon usage data. To our knowledge, this is the only study that has analyzed lecture capture usage in this way, and differentiates our work. We note that this does not remove all observer bias from the analysis as there are various factors that are picked by experimenters that help determine the groups (e.g., value of k, identification of days of the week as the principal attribute, use of five minutes as a cut-off for viewership). Nonetheless, this method does help us more easily identify natural clusters of like activity for analysis.

4.2. Conclusions

In this work we have described how a model of learners can be built through interaction data, how this model can be generalized, and how robustness of such a general model can be described. We applied these techniques in the area of lecture capture, and have demonstrated that there exist several interesting groups of learners, one of which has significantly higher performance indicators than the others. Finally, we take initial steps towards using this modelling technique as a predictive indicator, and describe the issues that arise when this is done.

This work prompts consideration of a number of questions, all of which form our future work, including:

1. Is the relationship between *High Activity* usage and performance results causal? We have taken the first steps towards answering this question by examining the incoming grades of learners and determining there is no statistically significant difference between groups on incoming grades, yet significant difference in course achievement appeared. However, we did use overall grade point average, and it may be more appropriate to only look at grades that are from disciplines related to the course. Further, our study was a field study, and learners had the opportunity to use various other tools, technologies, and methods to improve their learning. Finding participants for a more controlled study may better clarify the relationship between academic performance and usage, though such an experiment would likely be difficult.
2. Does the modelling technique stand up across different disciplines? We have begun initial steps to apply this technique to domains in the humanities, social sciences, and professional schools and, while we anticipate it will be appropriate we note that there may be subtleties that require domain knowledge such as identifying the appropriate value for k. This may also stand true for the modelling of courses of different size, complexity (e.g. senior undergraduate versus junior undergraduate), and pedagogical approach.
3. For what kinds of courses is the lecture capture model appropriate? Applying the abstract model described in §3.1 requires a full term course with a single midterm, and thus far we have only replicated the results within large second year undergraduate Chemistry and Biomolecules courses. More work is needed to see if this model is appropriate for courses of other disciplines, sizes, complexity, or pedagogical approaches. Our initial investigations to answer this question have been looking at large first year undergraduate STEM courses, and preliminary results using the error measurement techniques described in §3.2 suggest that some customization of the model to year of study may be required.

Despite these open questions, we have shown that there exists a significant positive correlation between performance and usage of lecture capture tools. Further, we shown that on-demand use of lecture capture for midterms (the *Just-In-Time* cluster) is both a popular

approach to using lecture capture technology (e.g. $n = 86$ in the *Chemistry 2011* cohort) and does not provide a statistically significant benefit to midterm, final, or overall evaluation.

References

Abt, G., & Barry, T. (2007). The quantitative effect of students using podcasts in a first year undergraduate exercise physiology module. *Bioscience Education e-Journal, 10*.

Baker, R., Harrison, J., Thornton, B., & Yates, R. (2008). An analysis of the effectiveness of podcasting as a supplemental instructional tool: a pilot study. *College Teaching Methods & Styles Journal, 4*(3), 49–54.

Brooks, C., Epp, C. D., Logan, G., & Greer, J. (2011). The who, what, when, and why of lecture capture. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge – LAK '11* (pp. 86–92). New York, New York, USA: ACM Press. URL http://dl.acm.org/citation.cfm?doid=2090116.2090128 http://www.cs.usask.ca/grads/cab938/TheWhoWhatWhenandWhyofLectureCapture.pdf.

Brooks, C., McKenzie, A., Meyer, D., Moormann, M., Rihtar, M., Rolf, R., et al. (2011). OpenCast Matterhorn 1.1. In *Proceedings of the 19th ACM international conference on Multimedia – MM '11* (pp. 703–706). New York, New York, USA: ACM Press. URL http://dl.acm.org/citation.cfm?doid=2072298.2072424.

Carle, A. C., Jaffee, D., & Miller, D. (2009). Engaging college science students and changing academic achievement with technology: a quasi-experimental preliminary investigation. *Computers & Education, 52*(2), 376–380.

Dasgupta, S. (2008). *The hardness of k-means clustering* (Tech. rep.). San Diego: Department of Computer Science and Engineering, University of California. URL http://charlotte.ucsd.edu/dasgupta/papers/kmeans.pdf.

Heilesen, S. B. (2010). What is the academic efficacy of podcasting? *Computers & Education, 55*, 1063–1068.

Hodges, C. B., Stackpole-Hodges, C. L., & Cox, K. M. (2008). Self-efficacy, self-regulation, and cognitive style as predictors of achievement with podcast instruction. *Journal of Educational Computing Research, 38*(2), 139–153.

Kurtz, B. L., Fenwick, J. B., Jr., & Ellsworth, C. C. (2007). Using podcasts and tablet pcs in computer science. In *Proceedings of the 45th annual southeast regional conference. ACM-SE 45* (pp. 484–489). New York, NY, USA: ACM. URL http://doi.acm.org/10.1145/1233341.1233428.

Leadbeater, W., Shuttleworth, T., Couperthwaite, J., & Nightingale, K. P. (2013). Evaluating the use and impact of lecture recording in undergraduates: evidence for distinct approaches by different groups of students. *Computers & Education, 61*(0), 185–192. URL http://www.sciencedirect.com/science/article/pii/S0360131512002163.

McCombs, S., & Liu, Y. (2007). The efficacy of podcasting technology in instructional delivery. *International Journal of Technology in Teaching and Learning, 2*(3), 123–134.

McKinney, D., Dyck, J. L., & Luber, E. S. (2009). itunes university and the classroom: can podcasts replace professors? *Computers & Education, 52*(3), 617–623. URL http://www.sciencedirect.com/science/article/pii/S036013150800167X.

Pursel, B., & Fang, H.-N. (2012). *Lecture capture: Current research and future directions*. The Schreyer Institute for Teaching Excellence, Pennsylvania State University (Tech. rep.).

Smith, G., & Fidge, C. (2008). *On the efficacy of prerecorded lectures for teaching introductory programming*. In *Proceedings of the tenth conference on Australasian computing education* (Vol. 78); (pp. 129–136). Wollongong, NSW: Australian Computer Society, Inc.

University of Purdue. (2012). Course signals – stoplights for student success. URL http://www.itap.purdue.edu/learning/tools/signals/ Accessed 28.10.2012.