Characterizing MOOC Learners from Survey Data Using Modeling and n-TARP Clustering

Mr. Taylor V. Williams, Purdue University, West Lafayette

Taylor Williams is a Ph.D. student in Purdue's school of engineering education. He is currently on an academic leave from his role as an instructor of engineering at Harding University. While at Harding he taught undergraduate biomedical, computer, and first-year engineering. Taylor also spent time working in industry as a systems engineer. Taylor received his master's in biomedical engineering from Tufts University and his bachelor's in computer engineering and mathematics from Harding University. His primary research interest is in how to use machine learning in fully online and hybrid educational environments to understand students and improve their learning.

Dr. Kerrie A. Douglas, Purdue University, West Lafayette

Dr. Douglas is an Assistant Professor in the Purdue School of Engineering Education. Her research is focused on methods of assessment for large-scale learning environments.

Mr. Tarun Yellamraju, Purdue University, West Lafayette

Tarun Yellamraju is currently a PhD student in the school of Electrical and Computer Engineering at Purdue University. He received his Bachelor of Technology with Honors degree in Electrical Engineering from the Indian Institute of Technology Bombay. His current research interests include High Dimensional Data Analysis and Machine Learning.

Prof. Mireille Boutin, Purdue University, West Lafayette

Mireille (Mimi) Boutin is an Associate Professor in Purdue's School of Electrical and Computer Engineering with a courtesy appointment in the Department of Mathematics. Her past research accomplishments include the development of light-weight methods for language translation on mobile phones, food analysis tools for the treatment of the inherited metabolic disease phenylketonuria, and improved document processing methods for the printing industry. Her current areas of research include signal processing, big data, and various applied mathematics problems motivated by engineering applications. In particular, she is interested in high-dimensional machine learning problems that stem from applications, including data analysis issues related to STEM education research. She created "Project Rhea," a student-driven online learning project at www.projectrhea.org. She is a three-time recipient of Purdue's Seed for Success Award. She is also a recipient of the Eta Kappa Nu Outstanding Faculty Award, the Eta Kappa Nu Outstanding Teaching Award and the Wilfred "Duke" Hesselberth Award for Teaching Excellence.

Characterizing MOOC Learners from Survey Data Using Modeling and n-TARP Clustering

Taylor V. Williams, Kerrie A. Douglas, Tarun Yellamraju, and Mireille Boutin
Purdue University

Abstract:
MOOCs (Massive Open Online Courses) attract a diverse and large set of learners, with largely unknown learning needs and expectations. Researchers have been exploring why learners enroll in MOOCs and have found that learners enroll for a variety of reasons. Knowing who these MOOC students are is an important step in improving their educational experience and the value of MOOCs. It is therefore vital to identify and understand what distinct student groups exist in a MOOC, that is, to learn who they are and what they want. Pre-course surveys try to collect this information by asking students about who they are and what they want from the MOOC they are enrolling in. However, making sense of this survey data is challenging.

Machine learning clustering techniques are a standard tool for identifying groups within data; however, two problems exist when trying to cluster survey data: (1) it is often not in a form easily interpreted by clustering algorithms and (2) survey data is frequently high dimensional, which standard clustering techniques cannot handle well. We describe a technique for converting survey data into machine interpretable feature vectors. We then propose analyzing the data using the n-TARP clustering technique which is capable of efficiently finding multiple different cluster solutions and is scalable to high dimensional data.

Using the proposed analysis approach on pre-course survey data from four MOOCs resulted in multiple distinguishable groups (i.e., clusters) of learners in each course, thus confirming the existence of many different survey response patterns. Additionally, these criteria persist between STEM and non-STEM courses. That is, we found learners grouped into similar clusters regardless of the course topic. The ability to separate learner types into distinct categories within and across courses is an important step in furthering the goal of enabling MOOC designers to design better online open educational systems that serve their diverse set of learners.

Keywords: clustering, MOOCs, survey analysis, open online education, educational data mining

1. INTRODUCTION

MOOCs (Massive Open Online Courses) often attract a diverse set of learners with a variety of needs and goals [1]. This diverse learner base provides an opportunity to learn about how and if learner goals vary by examining MOOC data. An understanding of these data can inform the design of future learning opportunities in engineering education and other fields. One way to identify these learner needs and goals is through MOOC pre-course surveys. However, because of the potentially vast number of learners in a MOOC, finding meaningful ways to utilize this survey data is challenging. If there exist a diverse set of learners within MOOCs, then simply looking at the learners' survey responses will not help us understand the people within those groups since important differences will be averaged out.

MOOCs are often designed with only a single type of user in mind—namely, someone who follows all the course material in a sequential manner. In other studies, these MOOC learners have been called "completing" [2] or "fully engaged" [3], but these learners are not the only type of people in the course. It is common for a large portion of the learners to either disengage with the course before completing it or to only access certain types of materials. Given MOOC learners' low completion rate, many researchers have argued that not all learners even intend to complete the course. Moreover, if some learners never intended to complete the course, then, completion is no longer a meaningful indicator of success for MOOCs [4]. Therefore, we need to develop other, more

meaningful metrics of student success and course evaluation. One such metric could come from improving our understanding of the variety of learner goals and intentions within these open online courses.

What is needed is a way to identify which individuals are similar, group them, and then characterize the diverse learner groups in useful ways. While there are many statistical techniques for finding patterns in data—including clustering data into distinct groups—these techniques typically require the data to be in the form of vectors in some space equipped with a distance measure. However, survey data is challenging to translate into this form, as answers can be missing or inaccurate, and survey questions may differ between courses or even from one course offering to the next. Additionally, survey results are likely to result in high-dimensional vectors characterizing each learner, and high-dimensional data is notoriously difficult to cluster. Furthermore, with diverse learners, there are likely numerous different statistically valid ways to group the learners into clusters, but most clustering techniques generate only a single cluster solution.

In this paper, we use a recent clustering technique on pre-course surveys data to identify learner groups with different motivations and intentions. Through pre-course surveys, learners can provide self-reported information such as their motivations, expected obstacles, hopes, or prior knowledge and experience. This learner-provided information hopefully contains useful clues on how to help learners achieve their goals, clues which could help inform the design of future learning environments according to the needs and goals of these different learner groups.

From an educational data mining perspective, analyzing the motivations of MOOC learners is a challenging task for two different reasons. First, survey answers must be represented in a consistent and relevant numerical way so that they can be analyzed efficiently. Second, the number and diversity of the learners suggest many different patterns of motivation may exist, each of which may be a reasonable way to group the learners. We address these two issues by extending the clustering approach of [5] which presents a computationally fast clustering method called n-TARP. Our extension is designed to find patterns in qualitative data coming from pre-course surveys. Briefly, the n-TARP works by generating and testing many randomly generated criteria for identifying groups (i.e., clusters) within the data. The feature vectors describing each student are constructed by classifying the survey questions into broader categories. Each learner's answers to the categorized questions are then modeled using a parametric random process.

In the following, we propose a n-TARP-based method to analyze pre-survey Likert scale questions to identify distinct subgroups of students. Using this method, we investigated the questions "Does the n-TARP clustering technique result in interpretable groups of MOOC learners?" and "If so, what can we learn about these learner groups?" To this end, we tested the n-TARP method on pre-course survey data acquired in multiple MOOCs offered on a popular MOOC platform. We examined four courses, two undergraduate STEM courses which would be required in their respective programs (an undergraduate mechanical engineering course and an applied science course) and two elective courses (a mathematics course and a course on wellbeing). These courses' pre-course surveys asked learners questions concerning their goals and applications for the course and their intended level of participation, among others. We hypothesize that MOOC learners form groups with distinct characteristics, and that these groups can be identified using pre-survey data.

2. RATIONALE FOR SELECTING CLUSTERING METHOD

Clustering is a well-studied unsupervised machine learning problem. Many popular clustering methods are available for clustering points in a low dimensional space: for example, k-means [6], [7], kernel k-means [8], the EM (Expectation Maximization) algorithm [9], and DBSCAN [10]. These methods are quite effective at identifying clusters of points in low dimensional space. However, in higher dimensional spaces the problem of finding clusters becomes more complicated due to the

curse of dimensionality [11]. The *curse of dimensionality* is the idea that data with many dimensions are inherently sparse—the data points are distant from each other—causing clustering algorithms to struggle to find similarities (and therefore clusters) between the data points. In the current study, dimensions are characteristics describing a person. For example, some dimensions of a person could include their age, where they were born, and their field of study. In this study, we examined 15 learner dimensions as indicated on a course pre-survey—dimensions related to their interests, the applications of what they hope to learn, and their perception of how their lifestyle may or may not inhibit their completion of this MOOC.

Several methods have been developed to address the complexity of finding patterns in high dimensional data (e.g., CLIQUE [12], FIRES [13], and PROCLUS [14]). However, typical approaches to high dimensional clustering are deterministic and seek to find only a single cluster structure solution within the data. Previous research has found that non-deterministic approaches can find multiple different cluster solutions, each which corresponds to a different separation criterion (separation criteria are used to determine cluster assignments) [15].

Previously, our team presented a clustering method called RP1D which can efficiently identify multiple cluster solutions, even within high dimensional data [5]. RP1D is a hierarchical clustering method, that is, a clustering method where each identified cluster is input back into the algorithm to identify further cluster subgroups. Each iteration in hierarchical clustering is called a level. In the current study, we used a single level adaptation of RP1D called n-TARP. In the next section, we summarize the n-TARP method as it was first introducted in [15]. n-TARP's single level implementation of RP1D was sufficient in this study as we were looking for multiple separation criteria to apply to the *whole* dataset. Our team has also previously shown that n-TARP is useful for finding multiple different patterns in the data, and that it is scalable to high-dimensions (i.e., greater than 40 dimensions) with only a modest increase in computational cost (as shown in [5]) while still yielding statistically significant clusters (demonstrated in [15]).

3. METHODS

Figure 1 shows the steps in our data modeling and clustering method for survey data, which we extended from the approach described in [15]. We first grouped similar questions so to model the learners' response patterns as a parametric random process. We then used the parameters of that random process to represent each learner as a vector in a high-dimensional space. We then clustered the learner vectors using n-TARP—the one level version of the random projection into 1D (RP1D) method of [5] (as done in [15]). We expected this approach to find multiple binary clusters. A binary cluster is a set of exactly two clusters into which the whole set of learners is split. Each of the expected multiple binary clusters is associated with a specific separation criterion which specifies how the learners are split into the two groups.

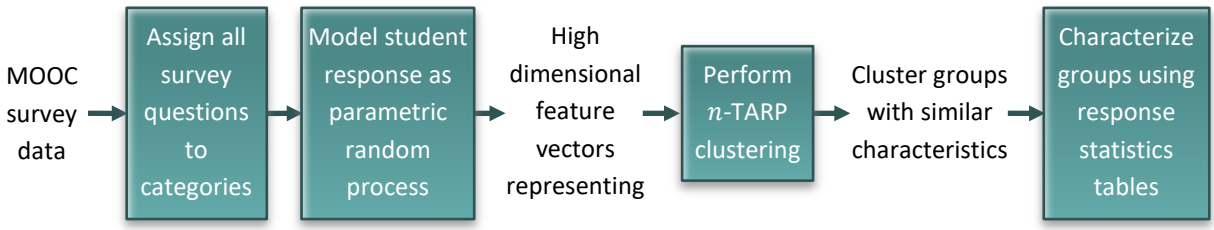


Figure 1: Flow diagram of the data analysis. Text outside a box describes the data (either raw or processed). Text within boxes describe actions performed on the data.

Modeling the data as a random process

As a preliminary test of the method, we coded eight 5-level rating-scale (Likert) survey questions into three categories nominally called C_1, C_2, C_3. We designated the rating scale levels as R_1 for

strongly agree to R_5 for *strongly disagree*. We viewed the set of possible combinations of question category, C_i, and response, R_j, as a random observation obtained by sampling among the categories and responses following some (unknown) probabilities. We estimated these probabilities by calculating the frequency of each respondent's category/response combination. In this way, each category/response combination becomes the estimated probability for a dimension describing that student. We stored the estimated probabilities for each student as the entries of a p-dimensional feature vector of real numbers which exists in the real-number space of $\mathbb{R}^p$ where p is the number of dimensions that ultimately described each student. The p-dimensional vector data thus represented each learner's probability of responding to each question code, C_i, with the response, R_j.

As an example, given three rating scale questions with five-levels ($R_1 - R_5$) all coded as C_1, a learner who responds *strongly agree* (R_1) to one of those questions and *agree* (R_2) to the other two questions has the following five-dimensional vector for questions categorized as C_1 (in this example $p = 5$ and the vector is in $\mathbb{R}^5$):

$\boldsymbol{C_1(R_1)}$	$\boldsymbol{C_1(R_2)}$	$\boldsymbol{C_1(R_3)}$	$\boldsymbol{C_1(R_4)}$	$\boldsymbol{C_1(R_5)}$
$\left(\frac{1}{3}\right) \cong .33$	$\left(\frac{2}{3}\right) \cong .67$	$\left(\frac{0}{3}\right) = 0$	$\left(\frac{0}{3}\right) = 0$	$\left(\frac{0}{3}\right) = 0$

We then store these five values as part of that student's feature vector. The feature vector can grow beyond five dimensions if other categories also exist for that student (for example, C_2). The feature vector could grow to many dimensions if it includes many categories and responses.

We recorded the feature vector for each student who fully completed the survey. For simplicity, we removed students with missing responses before proceeding to the next step—identifying clustering criteria.

Clustering

To identify clusters within these feature vectors, we used the n-TARP method [5]. The method has been shown to be effective for educational data in [15] even though it was initially developed for images [16].

The n-TARP clustering process as implemented is summarized below (from [5]):

1. Generate a random independent and identically distributed (i.i.d.) vector in p-dimensions where each element is drawn from a uniform random real-number distribution in the range [0, 1].
2. Project the feature vectors (representing the students' survey responses) onto the random vector (RV) by taking the dot product of the two vectors (Equation 1). Find the projection value for each student vector onto the random vector. The collection of all projections is designated the *set of projection values*.

$$\boldsymbol{a} \cdot \boldsymbol{b} = \sum_{k=1}^{p} a_1 b_1 + a_2 b_2 + \cdots + a_k b_k \qquad \textit{Equation 1}$$

   Where $\boldsymbol{a}$ is a student's feature vector, $\boldsymbol{b}$ is a random vector, and p is the number of dimensions present in the vectors.
3. (Optional) Construct a histogram of the set of projection values. If the histogram is bi-modal, then a successful pair of cluster groups has been found. This random vector is a separation criterion candidate.
4. For the set of projection values, search through the possible projection values for the threshold value, T, that minimizes W, the total intra-class variance renormalized by the variance of the projected data as discussed in [5], [16]. W measures to what extent the set of projected values divide into two groups. Small values of W are associated with a well-

defined binary clustering. We use the threshold value of $W = 0.36$ to select well clustered projections, as suggested by [5], [16]. G*roup 1* is defined as students whose projected value is smaller than the threshold, T, and *Group 2* as equal to or above T.

5. Repeat the above process steps n times (note, the total number of iterations is the n parameter in n-TARP).
6. Identify separation criteria as those random vectors that produce $W < 0.36$ with the data.

Grouping descriptions with statistical response pattern tables

Once we identify the clusters, we then average each group's responses in each dimension and present this data in a tabular form. We compare the average responses of each group to identify the differences between the two, i.e., the criteria that resulted in their separation. We can compare the group responses with the corresponding random vector entries to gain insights on the separation criterion. We will further elaborate on this analysis procedure while describing our experiments and results in the next section.

4. EMPIRICAL STUDY

We studied four courses offered by a large MOOC provider (listed in Table 1). We choose these courses for their variety of class sizes (119–3,275 students) and subject areas (within and outside of STEM) as well as their being either core and elective courses. In our analysis of these courses, we used all eight of the rating scale questions from the pre-course surveys. We list these eight rating scale questions in Table 2. We used the rating scale questions because of their similar form which expedited this exploratory study.

Table 1: Courses and post-cleaning student counts

Course number	Course subject area	Learner count: Available (Post-cleaning[3])
1	Applied science[1]	2,032 (1,866)
2	Mathematics[2]	750 (695)
3	Undergraduate mechanical engineering[1]	119 (100)
4	Wellbeing[2]	3,275 (2,987)

Notes: [1]Undergraduate core course, typical within its respective program; [2]Elective courses; [3]We removed learners with incomplete responses for this exploratory analysis

For compatibility with the n-TARP based method discussed in section 3, we grouped the eight questions into three categories. The questions and their category codes are listed in Table 2. The three categories in Table 2 ($C_1 \equiv$ Personal Interest, $C_2 \equiv$ University Application, and $C_3 \equiv$ Fit with Lifestyle) and the five possible response options to each questions ($R_1 - R_5$), means the resulting feature vectors belong in $\mathbb{R}^{15}$ (that is, three categories times five possible responses result in 15-dimensional vectors). While 15-dimensional data is only moderately high dimensional, in future work, we will extend our method to deal with additional survey questions which will increase the dimensionality of the data. The method at its core is easily scalable and designed for higher dimensional data.

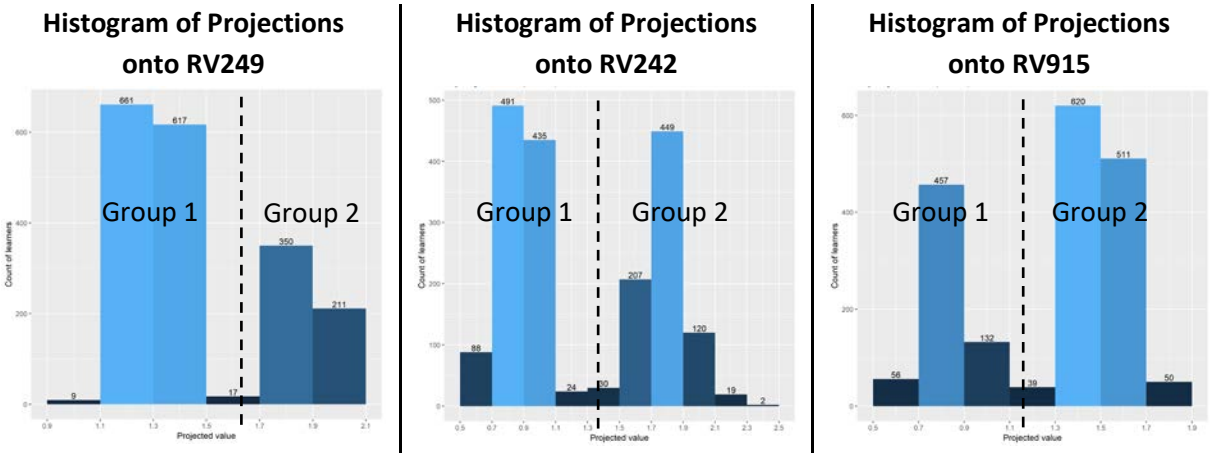
Table 2: Rating scale pre-survey questions analyzed. For each of these eight statements, the learners were asked "To what extent do you agree with the following statements?" and could select from a five-point rating scale ranging from "Strongly Agree" to "Strongly Disagree." The statements are shown here in the order asked.

Pre-course survey statement (i.e., question)	Code
"I'm taking this because I want to learn about the subject"	Personal Interest
"I'm taking this course to do my current job better"	Personal Interest
"I'm taking this course to improve my career prospects"	Personal Interest
"I'm taking this course to support current or future studies"	Personal Interest
"To support a university application"	University Application
"To find out more about the institution running the course"	Personal Interest
"I'm taking this course to find out how FutureLearn works"	Personal Interest
"I'm taking this course because I can fit it round my lifestyle"	Fit with Lifestyle

Results and discussion

Applying our clustering approach to our survey data with n = 1000 did, in fact, reveal multiple distinct learner groupings. This finding confirms our hypothesis that there exist multiple measurably distinct ways to group learners in these MOOCs. That is, there is more than one way to justifiably group a set of learners.

In each of the four courses examined in this preliminary study, we have found between 20–60 separation criteria (random vectors). Each of these criteria is used to split the learners into two distinct groups based only on their pre-survey responses. Figure 2 shows three representative histograms using the best three criteria identified in the applied science course. Notice the low W value and the well-separated bimodality seen in each histogram. Marked on the histograms is a dashed vertical line showing the optimum threshold, T—the point which best separates the learners into two distinct groups, as described in [5]. (The method to calculate T was described in section 3.)



$W = 0.179; T = 1.65$	$W = 0.180; T = 1.37$	$W = 0.199; T = 1.12$
1297 (70%) learners in Group 1	1056 (57%) learners in Group 1	648 (35%) learners in Group 1
568 (30%) learners in Group 2	809 (43%) learners in Group 2	1217 (65%) learners in Group 2

Figure 2: Histograms of the top three random vector (RV) separation criteria (those that best minimize the Within Sum of Square value, W, for the applied science course's pre-surveys (N = 1,865). A total of 31 criteria (of 1000) provide adequate group separation (that is, having $W < 0.36$ according to [5]*).*

It is interesting to note the different response patterns of the students in different groups and to compare them with those of the entire applied science class. First, Figure 3(b) (the middle column of

Figure 3) shows the mean values for the 15-dimensional feature vector for all students in Course 1. We will be comparing the groups identified by n-TARP with this "baseline profile" (Figure 3(b)). Figure 3 also shows the response statistics for the first two criteria shown in Figure 2 (i.e., *Group 1* and *Group 2* identified by applying the separation criteria RV249 and RV242). Note that while each of these separation criteria identifies distinct groups, the group characteristics are very different.

Course 1

		RV number / Group				
		RV249		No RV used	RV242	
Qual. Code	Response option	1	2	All	1	2
Personal interest	Strongly agree	28%	25%	27%	21%	35%
	Slightly agree	16%	15%	15%	12%	20%
	Neither agree nor disagree	22%	23%	22%	14%	33%
	Slightly disagree	9%	9%	9%	11%	5%
	Strongly disagree	26%	28%	26%	42%	6%
University application	Strongly agree	9%	9%	9%	0%	21%
	Slightly agree	8%	9%	8%	0%	19%
	Neither agree nor disagree	24%	29%	26%	0%	59%
	Slightly disagree	12%	11%	11%	20%	1%
	Strongly disagree	47%	42%	45%	80%	0%
Fit with lifestyle	Strongly agree	55%	0%	38%	40%	36%
	Slightly agree	45%	0%	31%	32%	30%
	Neither agree nor disagree	0%	65%	20%	15%	26%
	Slightly disagree	0%	13%	4%	4%	4%
	Strongly disagree	0%	22%	7%	9%	3%

0% 100%

(a) (b) (c)

Figure 3: For course 1's top two separation criteria (RV249 and RV242 shown in (a) *and* (c)*, respectively), the response pattern statistics for the applied science course result in distinct response groups (labeled Group 1 and Group 2, matching the labels from Figure 2). The dimensions that are unaffected by the criteria (i.e., personal interest and university application for RV249;* fit with lifestyle *for RV242) remain consistent (within 5%) of the mean responses for the entire set of learners (shown in* (b)*).*

When examining criterion RV249 we see that for each of the five response options, the *personal interest* and *university application* codes are very similar across all three groups (Group 1, Group 2, and all learners), differing by no more than 5%. For RV249 the *fit with lifestyle* code shows substantial distinctions between the two groups with practically all learners who agreed with the statement belonging to Group 1 and all who were neutral or disagreed in Group 2. When examining the question associated with the *fit with lifestyle* code—"I'm taking this course because I can fit it round my lifestyle"—we can see that criterion RV249 has effectively separated learners who believed they could fit this MOOC into their life (Group 1, containing 70% of the learners) from those who responded as unsure or negative about their MOOC-lifestyle fit (Group 2, containing 30% of the learners) without regard to their *personal interest* responses. It is interesting that so many learners (30%) enroll in a MOOC without believing they have time for it.

Looking at the groups found with criterion RV242 we see that the responses to the *fit with lifestyle* question are similar between the two groups; however, the *university application* responses are very well separated between groups. The *university application* code also contained only a single statement—"To support a university application"—with which the learner was asked to agree or disagree. One interpretation of the two groups is that those who agreed (Group 2, containing 43% of the learners) may be university researchers in the field of this applied science while the remaining learners have no direct university association. Criterion RV242 also shows some separation for *personal interest*, but this separation is not as pronounced as that of that for a *university application*. As a reminder, the pre-course survey questions associated with these codes are available in Table 2.

Persistence of patterns in other courses

Considering the patterns, we found (described above), we next investigated if the same patterns exist in the other courses in our study. In other words, we investigated if using the separation criteria found for course 1 (e.g., RV249 and RV242) on courses 2–4 also resulted in significant groupings of students. We repeated the process by using n-TARP to identify separation criteria in courses 2, 3, and 4 and then seeing if these identified criteria also identify significant groupings in the other 3 courses. We were interested to see if criteria from one type of course (e.g., a small undergraduate STEM course) remained valid when applied to other course types (e.g., non-STEM or elective or large courses). We expected different types of courses to attract different types of learners, so the criteria identified in one course may not result in distinct learner groups in a different type of course.

We show the results of this experiment in Figure 4. Several separation criteria do transfer successfully, identifying significant groupings in other courses. For example, when we use criteria from course 3 on courses 1, 2, and 4, a strong majority (60–80%) of the criteria are significant in the other 3 courses. On the other hand, we also have situations where the criteria are not highly transferable to at least one other course. For example, course 2's criteria do not transfer well to course 3 (less than 33% criteria transfer). The other extreme is also observed, as 100% of the criteria found using course 1 are valid for course 2. It is interesting to note that when using criteria from course 4, fewer than 50% of the criteria transfer to any other course. One possible explanation why course 4's criteria might have failed to separate students in the other courses is because course 4 (the wellbeing course) is non-STEM, unlike the remaining three courses. The observed lowered criteria transfer rate reaffirms the expectation that learners who enroll in STEM courses differ somewhat from those in non-STEM courses. However, it is still interesting to observe that a non-trivial number of criteria—even if not always a majority—do still transfer to the other STEM courses.

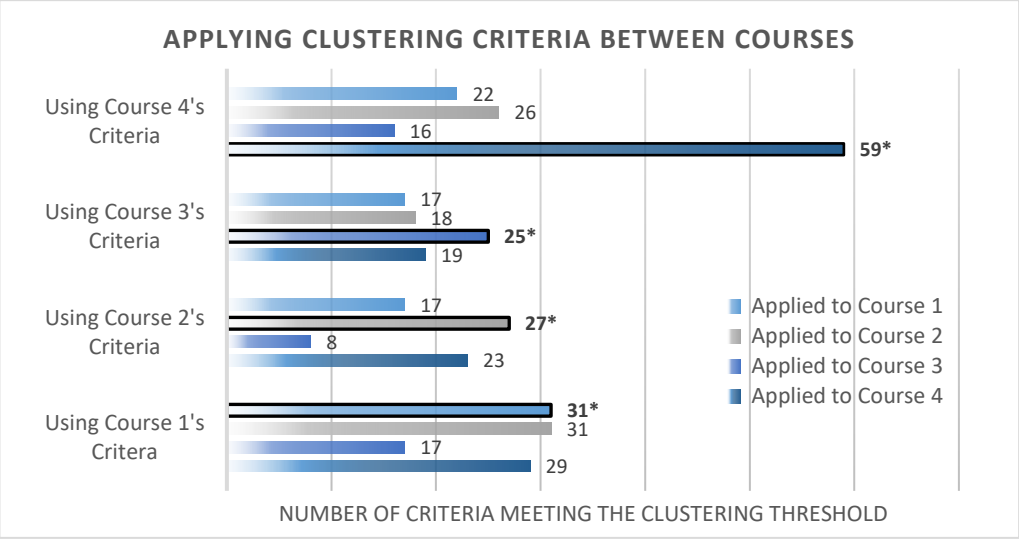
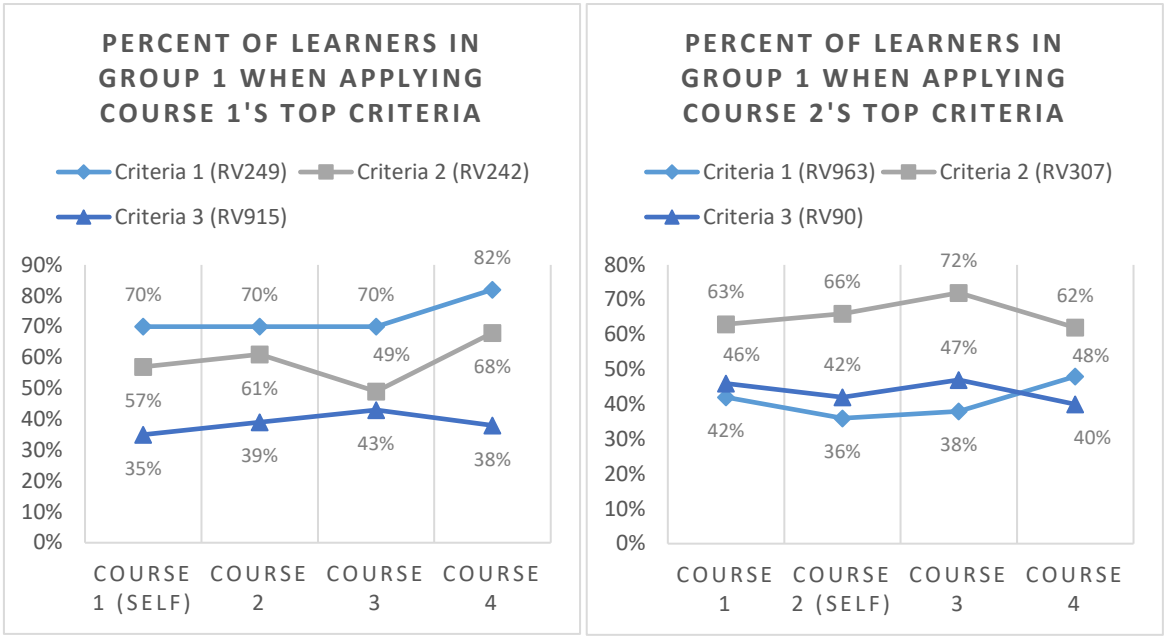


Figure 4: When applying criteria identified in one course to another course many of those criteria still meet the $W < 0.36$ threshold. In each group, the course from which the criteria were first identified is marked (*).

We next checked the student percentage split between groups using each set of top criteria from two courses (Course 1 and Course 2). To check if the split remained constant or not a single criterion from one course was applied to all four courses, then we found the proportion of learners in each resulting group (Figure 5). The group proportions remain remarkably similar across courses with a fluctuation of no more than 19% (the average fluctuation is 11.3%). This result further indicates that some of the criteria we identified in one course can identify similar groupings in other courses. It is

also indicative that the students' responses may be modeled by some global distribution that is independent of course.



(a) (b)

Figure 5: When applying the top three criteria from one course to the other three courses we find that a roughly consistent percentage of students fall into each group. Here we see the percentage of learners who belong to Group 1 when the top three criteria from Course 1 (a) and Course 2 (b) are applied to the other three courses. The remaining percentage of students are part of Group 2. In both cases, all three criteria result in measurably distinct groups in the other courses (that is, $W < .36$).

For the final data visualization, shown in Figure 6, we considered how well a single criterion found in one course applies to all four courses. That is, does a criterion identified in one course identify similar student groups when applied to other courses? We observe that the groups formed in each of the four courses are very similar. RV249 continues to separate the data based on *fit with lifestyle* (Figure 6(a)) while RV242 continues to separate based on *university applications* (Figure 6(b)). That is these criteria from course 1 do find similar patterns in the three other courses. Notice that Figure 6 also shows the mean values for all students in each course alongside each course's *Group 1* and *Group 2.* For Course 1 these columns contain the same values as previously shown in Figure 3(b).

Cross-Course Comparison of Mean Probabilities of Learners' Responses to Differently Coded Survey Questions

Groups identified using Course 1's: **RV249** **RV242**

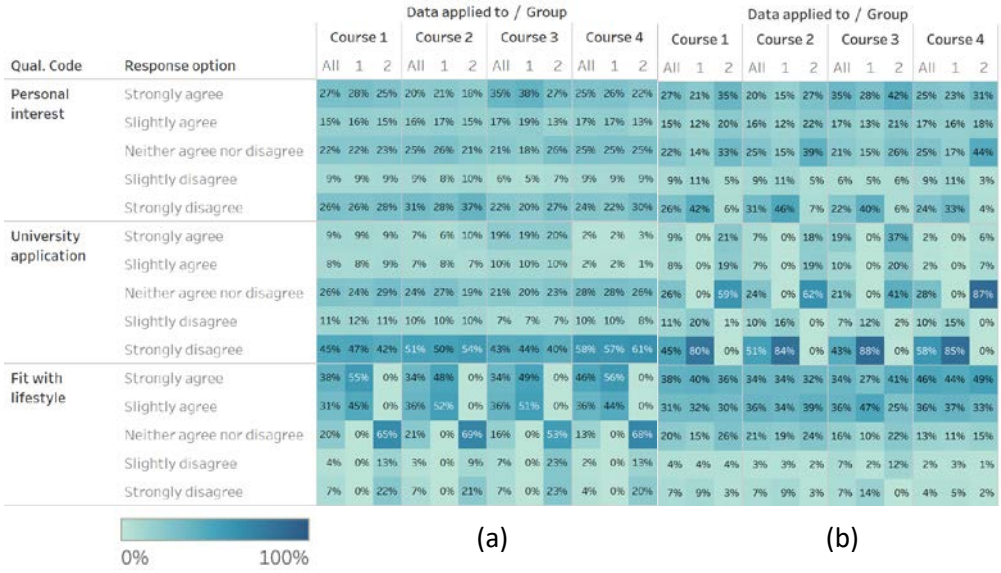


Figure 6: Applying Course 1's top two criteria (RV249 and RV242) to all four courses show that criterion RV249 differentiates learners based on their fit with lifestyle *responses (and not* personal interest *or* university application*) whereas RV242 differentiates on* personal interest *and* university application *(and not* fit with lifestyle*). Shown here are the observed probabilities for each response option by each course's Group 1, Group 2, and its entire set of learners ("All").*

Limitations

Some limitations of this study include that the dataset was pre-existing and provided as-is, so there was limited freedom in our choosing what questions we could analyze. Likewise, since we did not write the survey questions, we had to use our best judgment to determine how to group them. Also, due to the limited available questions we estimated some of the parameters in this study from a single question (i.e., *university applications* and *fit with lifestyle*); however, high-quality estimates for the random parameters of the learners' feature vectors require a lot of data, or, in this case, multiple questions in the same category to provide redundancy. Additionally, this study examined the learners from only a few courses—a small sample in comparison to the whole population space.

A non-critical mistake occurred during this study's n-TARP implementation, namely the random vectors were sampled from the real number range [0, 1] where they should have come from the range [-1, 1]. The implication of this mistake is that vectors from only a single quadrant were tested. It should be noted that within the other three quadrants there may exist more (and possibly better) criteria. Since we did successfully find separation criteria in this quadrant this issue is minimal. However, if we had not found patterns then we would have needed to explore the larger space (i.e., [-1, 1]) for successful criteria.

5. IMPLICATIONS AND FUTURE WORK

Implications

This study has successfully found significant groupings in student response data in a random manner. We have highlighted the best groupings and demonstrated that there exist distinct clusters of students independent of courses as well as specific to some courses. This difference in characteristics could be used to both better design common course features and more specific characteristics of specific courses to increase student engagement and productivity.

n-TARP has the additional benefit of being computationally inexpensive and scalable to high-dimensional data [5]. It is also non-deterministic which yields a distribution of clusters, each of which could identify different patterns in the data and yield distinct insights. The high-dimensional application of n-TARP will be especially useful when more conventional clustering approaches fail.

Future work

The work presented here is still in a preliminary stage. While the data we analyzed in this paper was only 15-dimensional, n-TARP can scale to high-dimensions [5] meaning that as the dimensionality of the survey data input increases n-TARP should continue to be effective. We intend to expand the dimensionality of the data by including all the questions from the pre-survey instead of the eight question subset we used in this study. This goal of including additional questions poses some challenges since not all questions have a Likert scale, nor are they all ordinal. Further, in this study we considered pre-course survey responses from only four courses; moving forward we plan to increase this analysis to over 200 courses.

6. CONCLUSIONS

In summary, we proposed a new data analysis approach for survey data using n-TARP. We generated a distribution of clusters based on student responses to MOOC pre-course survey questions. Our approach successfully found multiple significant ways to group (cluster) students based on their responses. We also observed that some separation criteria are specific to the course in which they are found whereas some other criteria can identify student groups in all the courses tested. These criteria, which identify different types of learners, could yield insights to help improve specific courses as well as improve the entire learning platform. Knowing to which of these n-TARP-identified groups a learner belongs may also aid in designing more effective learning environments for these individuals according to their needs and goals, ultimately improving students' online learning experience.

7. ACKNOWLEDGMENTS

This work was made possible by the U.S. National Science Foundation (NSF) (PRIME #1544259). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF.

The authors would like to thank FutureLearn for providing the data and the many reviewers who made this a much stronger paper.

8. REFERENCES

[1] R. F. Kizilcec and C. Brooks, "Diverse big data and randomized field experiments in MOOCs," in *Handbook of Learning Analytics*, 1st ed., C. Lang, G. Siemens, A. Wise, and D. Gasevic, Eds. Society for Learning Analytics Research (SoLAR), 2017, pp. 211–222.

[2] R. F. Kizilcec, C. Piech, and E. Schneider, "Deconstructing disengagement: analyzing learner subpopulations in massive open online courses," in *Proceedings of the Third International Conference on Learning Analytics and Knowledge - LAK '13*, 2013, pp. 170–179.

[3] K. A. Douglas, P. Bermel, M. M. Alam, and K. Madhavan, "Big data characterization of learner behaviour in a highly technical MOOC engineering course," *J. Learn. Anal.*, vol. 3, no. 3, pp. 170–192, 2016.

[4] D. Koller, A. Ng, and Z. Chen, "Retention and intention in massive open online courses: In depth," *Educ. Rev.*, vol. 48, no. 3, pp. 62–63, 2013.

[5] T. Yellamraju and M. Boutin, "Clusterability and clustering of images and other 'real' high-dimensional data," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1927–1938, 2018.

[6] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *J. R. Stat. Soc. Ser. C (Applied Stat.*, vol. 28, no. 1, pp. 100–108, 1979.

[7] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, 2010.

[8] B. Schölkopf, A. Smola, and K.-R. Müller, “Nonlinear component analysis as a kernel eigenvalue problem,” *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, Jul. 1998.
[9] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *J. R. Stat. Soc. Ser. B*, vol. 39, no. 1, pp. 1–38, 1977.
[10] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in *Proceedings KDD*, 1996, vol. 96, pp. 226–231.
[11] R. Xu and D. Wunsch II, “Survey of clustering algorithms,” *IEEE Trans. Neural Networks*, vol. 16, no. 3, pp. 645–678, 2005.
[12] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, “Automatic subspace clustering of high dimensional data for data mining applications,” in *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data - SIGMOD ’98*, 1998, pp. 94–105.
[13] H.-P. Kriegel, P. Kroger, M. Renz, and S. Wurst, “A generic framework for efficient subspace clustering of high-dimensional data,” in *Fifth IEEE International Conference on Data Mining (ICDM’05)*, 2005, no. 5, pp. 250–257.
[14] C. C. Aggarwal, J. L. Wolf, P. S. Yu, C. Procopiuc, and J. S. Park, “Fast algorithms for projected clustering,” in *SIGMOD ’99 Proceedings of the 1999 ACM SIGMOD international conference on Management of data*, 1999, vol. 28, no. 2, pp. 61–72.
[15] T. Yellamraju, A. J. Magana, and M. Boutin, “Investigating students’ habits of mind in a course on digital signal processing,” *in-press*, pp. 1–8, 2018.
[16] S. Han and M. Boutin, “The hidden structure of image datasets,” in *2015 IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 1095–1099.