JOURNAL OF LEARNING ANALYTICS
SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH
Volume 10(1), 113–132. https://doi.org/10.18608/jla.2023.7807

The Promise of MOOCs Revisited? Demographics of Learners Preparing for University

Michael J. Meaney[1], Tom Fikes[2]

Abstract
This paper leverages cluster analysis to provide insight into how traditionally underrepresented learners engage with entry-level massive open online courses (MOOCs) intended to lower the barrier to university enrolment, produced by a major research university in the United States. From an initial sample of 260,239 learners, we cluster analyze a subset of data from 29,083 participants who submitted an assignment in one of nine entry-level MOOC courses. Manhattan distance and Gower distance measures are computed based on engagement, achievement, and demographic data. To our knowledge, this marks one of the first such uses of Gower distance to cluster mixed-variable data to explore fairness and equity in the MOOC literature. The clusters are derived from CLARA and PAM algorithms, enriched by demographic data, with a particular focus on education level, as well as approximated socioeconomic status (SES) for a smaller subset of learners. Results indicate that learners without a college degree are more likely to be high-performing compared to college-educated learners. Learners from lower SES backgrounds are just as likely to be successful as learners from middle and higher SES backgrounds. While MOOCs have struggled to improve access to learning, more fair and equitable outcomes for traditionally underrepresented learners are possible.

Notes for Practice

- For the first time, cluster analysis is used to explore behaviour patterns of traditionally underrepresented learners engaging in entry-level MOOC courses produced by a major research university explicitly seeking to lower barriers to entry for university matriculation.
- Learners without a college degree are more likely to be high-performing compared to college-educated learners.
- For a subset of data for which approximated SES data was available, learners from lower SES backgrounds are just as likely to be successful as those from middle and higher SES ones.
- Utilizing a categorical variable like education level to derive clusters using the Gower distance metric allows for mixed-variable clustering, and may reveal subgroups of underrepresented learners potentially more amenable to support and remediation.
- Results in this paper suggest that taking a more explicit approach to analyzing MOOC data across demographic variables would be a promising way to move MOOCs back toward their original mission of widening educational access.

Keywords
Distance education, massive open online courses, MOOCs, learning analytics, learning design

Submitted: 01/07/2022 — **Accepted:** 23/01/2023 — **Published:** 09/03/2023

Corresponding author [1]Email: mmeaney@asu.edu University of Cambridge and Arizona State University. Address: EdPlus Action Lab at Arizona State University, 1365 N Scottsdale Rd, Scottsdale, AZ 85257, USA. ORCID ID: https://orcid.org/0000-0002-4497-0362
[2]Email: thomasfikes@me.com Address: EdPlus Action Lab at Arizona State University, 1365 N Scottsdale Rd, Scottsdale, AZ 85257, USA. ORCID ID: https://orcid.org/0000-0001-7060-3913

1. Introduction

This JLA special issue on Fairness, Equity, and Responsibility aims to augment the existing literature by expanding converge of "the design, implementation, risks and benefits of learning analytics for minority and disadvantaged groups, such as: students with disability, subcategories of unexplored demographic categories, migrants, etc."1 This paper contributes to this

[1] https://learning-analytics.info/index.php/JLA/announcement/view/177

aim by exploring how learners from traditionally underrepresented backgrounds among MOOC users — students without a college degree, and students from lower socioeconomic status (SES) backgrounds — are utilizing MOOCs offered by a research-intensive university in the USA that are intentionally designed to be more inclusive. It does so by utilizing a common learning analytic technique, cluster analysis, on data from entry-level MOOC courses explicitly designed to lower barriers to entry for university matriculation. To date, while the learning analytic literature has robust coverage of clustering techniques on MOOC data (Li & Baker, 2018; Ferguson & Clow, 2015; Kizilcec et al., 2013), these techniques are not used to explore questions of fairness and equity for underrepresented leaners (Meaney, 2021). Furthermore, no literature examines these questions in the context of entry-level MOOCs produced by American research-intensive universities intending to lower barriers to university enrolment. Additionally, a mixed-variable clustering approach is taken in one analysis utilizing Gower distance measures, which enables clusters to be formed by both demographic and performance data (Meaney & Fikes, 2022). This is one of the first instances that such an approach has been utilized in the MOOC literature; it provides an important template to consider when exploring questions of fairness and equity because it enables categorical variables like demographics to be factored directly into cluster analysis.

In the first set of analyses, clusters based on participation/performance data from 29,083 learners are enriched by considering how learners with different educational backgrounds are dispersed throughout. A second, less common method of cluster analysis is then used, which explicitly considers education background in computing distance measures. Additionally, data on median household income from the 2016 American Community Survey, at the Census tract level, is matched to a smaller subset of 2,343 learners to further investigate the clustering profiles of users based on approximated SES.

The first clustering approach yielded the same four to five clusters commonly observed in the research literature (Li & Baker, 2018; Kizilcec et al., 2013). Learners without a college degree, however, were found more likely to be successful in the set of MOOCs analyzed compared to their better-educated peers; additionally, in the second clustering approach, learners from low-SES backgrounds were found to perform no worse than peers from high-SES backgrounds.

In this paper, MOOCs refer to Coursera- and edX-style xMOOCs produced in the USA, predominantly in English, and which stipulate open enrolment without entry qualifications. These MOOCs have no barriers to access content (though the content may be copyrighted and thus not meet the "open" definition of OER), are online and are available to anybody with an internet connection, and are free to complete though they may charge a fee for certification (Deng et al., 2019).

2. Literature Review and Research Questions: MOOC Learners and their Characteristics

Investigating behaviour patterns of MOOC users is central to the existing learning analytics literature (Joksimović et al., 2019; Gardner & Brooks, 2018). Researchers have documented extensively the extent to which MOOCs reflect first-order digital divide issues around gaps in access to and utilization of technology (Reich & Ruipérez-Valiente, 2019; Rohs & Ganz, 2015). Additionally, cluster analysis and other computational methods have been used to understand different behaviour patterns of subgroups within MOOCs (Li & Baker, 2018; Ferguson & Clow, 2015; Kizilcec et al., 2013). The literature has yet to explore, however, whether behavioural subgroups differ across demographic characteristics, particularly those revealing dimensions of underrepresented status, like educational background and SES; this paper seeks to fill this gap.

We do, however, know quite a bit of other information about MOOC users. While MOOC participants are heterogenous and demographic data on them is limited, some common characteristics are frequently observed. Most users are located outside of the USA and hold a bachelor's degree (Reich & Ruipérez-Valiente, 2019; Rohs & Ganz, 2015). Some 90% of enrollees do not complete their course (Jordan, 2014).

Beyond categorization, researchers have analyzed the specific behaviour patterns of MOOC learners. Kizilcec et al. (2013) use K-means to cluster analyze more than 90,000 learners enrolled in three computer science MOOCs. They determine four prototypical user types: completing, auditing, disengaging, and sampling, defined by steep drop-out points and deeply unequal levels of participation. The methods employed by Kizilcec et al. (2013) have been leveraged across several different MOOC environments (e.g., Arora et al., 2017; Kovanović et al., 2016; Ferguson & Clow, 2015). With few exceptions (e.g., Ferguson & Clow, 2015), these papers consistently report four types of engagement patterns; this is further confirmed in the literature beyond research using strictly cluster analysis (Li & Baker, 2018; Anderson et al., 2014; Ramesh et al., 2014). As described by Li and Baker (2018), these engagement groups are sometimes labelled with different names, and sometimes a group may be split into a fifth group or combined into a third, but the prototypical patterns remain the same, and broadly speaking can be classified as follows: "disengagers," "auditors," "quiz-takers," and "all-rounders." "Disengagers" enrol in the course but have very little engagement thereafter; these students typically made up the largest subgroup. Auditors engage with the course material, but rarely submit assignments. "Quiz-takers" engage less with the course materials and content, but complete and submit assignments. All-rounders are students who engage most similarly to conscientious students in traditional courses, with high levels of interaction with course materials and assignment submissions (Li & Baker, 2018).

Though a robust literature has developed around analyzing student behaviour in MOOCs, progress in understanding questions about behaviour patterns potentially differentiated along demographic lines is more limited, especially regarding characteristics like educational background and SES (Joksimović et al., 2018; Deng et al., 2017). For the field to progress on its coverage of fairness and equity, this gap should be rectified.

2.1. What Do We Know about Underrepresented Learners in MOOCs?

The evidence that does exist on the relationship between education level and MOOC participation is mixed (Joksimović et al., 2018), with some researchers finding no relationship (Brooks et al., 2015), and others finding that more highly educated learners are more likely to persist (Kizilcec & Halawa, 2015).

Engle et al. (2015) find that completion and certification rates are positively associated with educational background, with more highly educated students more likely to pass a course and earn distinction. This replicates a finding of Guo and Reinecke (2014) who found that nearly 70% of students who completed a certificate held a bachelor's or master's degree. Greene et al. (2015) similarly found that higher levels of previous education were associated with less dropout. Others have found no association between prior level of educational attainment and outcome. Zhang et al. (2019) found no significant effect of students' educational background on course completion. Similarly, Goldberg et al. (2015) found no relationship between prior educational attainment and student outcomes.

Even less work has been done to understand the relationship between SES and MOOC participation. This is partly due to the difficulty of collecting SES data in the context of MOOCs. In general, the findings suggest that success in MOOCs is biased toward those from high-SES backgrounds, but that there may be interesting differences in motivations for enrolment between high-SES and low-SES groups. Hansen and Reich (2015) found that the average MOOC user was more likely to live in wealthier and more highly educated neighbourhoods. Ganelin and Chuang (2019) analyzed user registrations for edX courses between 2012 and 2018, identifying the location of enrollees by both geolocation and user-reported mailing address, and found that registration rates are higher among more prosperous postal codes. Despite MOOCs mostly serving learners from more educated and wealthy backgrounds, there is some evidence that learners from more disadvantaged backgrounds are using MOOCs to seek professional advancement (Stich & Reeves, 2017; Zhenghao et al., 2015; Dillahunt et al., 2014).

It is worth noting that there is some coverage in the learning analytics literature of learner behaviour patterns among subgroups of non-traditional learners in more formal online learning programs. Recent work by Zamecnik et al. (2022) uses soft clustering techniques on engagement and survey data to develop learner profiles among 1919 students enrolled in an online degree program offered by an Australian university. They find that learners from a lower SES background are likely to enrol in Marketing and Law programs, and unlikely to enrol in Commerce and Science programs. This data set, analyzed by Zamecnik et al. (2022), is subject to the selection effect of students enrolling in formal online degree programs, and contains a relatively smaller sample size, making it qualitatively different than data from MOOC platforms.

Importantly, MOOCs, in the context of this paper, are courses that stipulate open enrolment without entry qualifications, have no barriers to access content, are online and available to anybody with an internet connection, and are free to complete though they may charge a fee for certification (Deng et al., 2019). These courses attract a wider range of students engaging in learning for a broader array of reasons (Gardner & Brooks, 2018; Kizilcec & Schneider, 2015).

2.2. Summary of the Literature and Research Questions

The literature is ambiguous regarding the relationship between student demographic characteristics and MOOC engagement and outcomes (Joksimović et al., 2018). At the same time, researchers have leveraged cluster analysis to explore different engagement and achievement patterns in MOOCs (Ferguson & Clow, 2015; Kizilcec et al., 2013). Furthermore, there is a need to better understand non-mainstream users of MOOCs if these courses are to broaden access to higher education for traditionally underrepresented learners, as was originally conceived (Deng et al., 2017). Concurrently, it is also unclear whether different subgroups of students display heterogenous engagement and achievement patterns (Gardner & Brooks, 2018). Based on this reading of the literature, we sought to replicate existing versions of cluster analyses and explore demographic subgroups within these clusters with a particular focus on underrepresented learners, and to do so in a context of explicitly entry-level MOOCs designed to lower barriers to entry for university matriculation. Doing so will expand the coverage in the learning analytics literature of questions of fairness and equity in scalable learning environments for traditionally underrepresented learners. We thus crafted the research questions as follows:

RQ1: Do learners in entry-level tertiary MOOCs demonstrate similar patterns of clustering found in the broader MOOC literature?

RQ2: Are demographic subgroups of learners, specifically along the educational background dimension, represented equally across clusters?

RQ3: What demographic and engagement insights can be unveiled through leveraging a more novel, demographically sensitive cluster analysis method?

3. Research Context, Data, and Methods

We maintained a formal academic appointment with a major research university in the USA, which provided research facilities, support, and access to data. While categorized as an R1, "Very high research activity" university according to The Carnegie Classifications of Institutions of Higher Education (2017), the university also maintains a commitment to inclusive higher education, with a particular focus on broadening access and success to underrepresented populations. Toward that end, the university developed a series of nine entry-level MOOCs that could help earn a student admission into the university. To adhere to a high level of privacy and duty of care for our research participants, as well as sensitive partnerships, we choose to not disclose the institution. This study was approved by the research ethics processes (IRB) of two major research universities, the host university, and the corresponding author's doctorate-granting university.

3.1. Data

Data for the analysis is derived from the host university's sequence of nine MOOCs that collectively represent entry-level university credits for first-year students. These courses offer a mix of science- and math-focused content alongside humanities and social science content. Specific courses themselves were not utilized as units of analysis to maintain focus on general patterns of engagement among underrepresented learners in enrolment and participation across courses. Data were collected between fall 2017 and spring 2020.

Data from a variety of different sources from the host university were merged to build the final data set, including edX enrolment data containing demographic information, host university registrar data, before the course survey data, host university gradebook data, edX activity log data, and American Community Survey (ACS) data (Chetty et al., 2017). Data on median household income from the 2016 ACS, at the Census tract level, is matched to a subset of users to further investigate their clustering profiles based on approximated SES.

Data cleaning and analyses were completed in R (R Core Team, 2019) utilizing the following programming packages: "distances" (Savje, 2021), "cluster" (Maechler et al., 2022), "nbclust" (Charrad et al., 2014), "factoextra" (Kassambara & Mundt, 2016), "tidyverse" (Wickham, 2022), and "data.table" (Dowle et al., 2022). The key variables of interest for the analysis were education level, percent grade, a computed participation/performance metric, event count, and SES. Each variable is described in turn.

Education Level: Education level was obtained from the edX enrolment data. Users could select from nine potential levels, including the following: None, Junior High School, High School, Postsecondary Degree, Associate's Degree, Bachelor's Degree, Master's Degree, Professional Degrees, and Other. Variables were recoded into College Plus, No College, and Unknown. Postsecondary degrees, including Associate's, Bachelor's, Master's, and Professional Degrees, were grouped into College Plus, while None, Junior High School, and High School were grouped into No College. Other and non-responses were grouped into Unknown.

Percent Grade: Percent grade represented the learners' total grade in the course, primarily comprised of quizzes and tests.

Participation/Performance Metric: Kizilcec et al. (2013) compute a description for each learner based on how that learner engaged throughout the course. They labelled participants during each assessment period as on-track (T; did the assessment on time), behind (B; completed the assessment late), auditing (A; watched videos and engaged materials but did not complete the assessment), or out (O; no course activity at all). Scores 0–3 were assigned to each participant (0 = O, 1 = A, 2 = B, 3 = T). Similarity scores were then computed for each participant using the Manhattan distance method, and then the K-means clustering algorithm was applied. Ferguson & Clow (2015) leveraged an augmented version of the Kizilcec et al. (2013) methodology on three FutureLearn courses. To provide a more granular accounting of learner behaviour and account for a wider range of actions, they coded student behaviour along an eleven-point framework, as opposed to a 0–3 scale.

We employ a method similar to Kizilcec et al. (2013) and Ferguson & Clow (2015) in computing the participation/performance score, with two important differences. First, we assign scores of 0–4, accounting for 0 = dropped from course, 1 = never-graded lurking (students who never submitted an assignment but engaged with course content), 2 = ever-graded lurking (students who submitted assignments but not in the relevant assessment period), 3 = graded and behind (students who submitted an assignment in the relevant assessment period but scored less than 70%), and 4 = graded and on track (students who submitted an assignment for the relevant assessment period and scored 70% or above). "Behind" in this context indicates that a student had not submitted an assignment scoring 70% or more for the relevant assessment period, not behind temporally. Second, we assigned this grade for each graded course sequential, rather than each assessment period, which ranges from course to course between 22 and 48 sequentials per course. This captures activities in between assessment periods, allowing for a more granular accounting of student participation/performance. To norm the score across the different course lengths, student total scores were summed and then divided by the number of total graded sequentials in the course.

Event Count: Total event counts are not included as features during the cluster analysis. Event counts, however, are considered in describing the clusters. Event count is a generic, numeric variable representing the total number of actions a learner executed during a course. This includes logins, video events, discussion forum events, events to access the syllabus

and reference material, and all click-based events a student can take. Activity-based features, like event counts, are commonly used in the literature as a measure of engagement (Gardner & Brooks, 2018). In this study, event counts are considered in aggregate and not differentiated between event type, as this level of data granularity was not available for our analysis.

SES Status: Survey data, while integrated into the final data set, was not utilized in the final analysis. The survey data did, however, provide a learner's longitude and latitude coordinates, which were paired with income-level data at the Census tract level to derive an approximated SES. This approach has been utilized before (Ganelin & Chuang, 2019) as well as similar approaches (Hansen & Reich, 2015). Only students from the USA who completed the before-course survey were assigned an SES, so it was a small subset compared to the entire sample. SES data is analyzed in the second clustering analysis.

SES was defined in relation to what is considered a "low-wage" worker: a person making two-thirds or less of the median national income (Escobari et al., 2019). Census tracts in which the median household income was two-thirds or below the national median income in 2016, calculated to be $41,609.68, were labelled low-SES, whereas all others were labelled medium- to high-SES.

Table 1 provides the enrolment data from the edX system on educational background level, gender, and geographic location for 260,239 learners who enrolled in the nine courses represented in the data, administered across 58 different course-terms. This represents the initial data set of users. College-educated learners are overrepresented, comprising 56% of the enrolment, compared to 23% of students without a college degree. While college-educated learners are overrepresented, the proportion is lower than commonly found in the MOOC literature, where college-educated learners typically comprise between 60–80% of enrolments (Meaney & Fikes, 2019).

Table 1: Descriptive Statistics for Enrolled Learners in the Study Sample

Characteristic	**Overall**, N = 260239	**USA**, N = 49195[†]	**Outside USA**, N = 179155[†]	**Unknown**, N = 31889[†]
Education_Level				
College Plus	145399 (56%)	20299 (41%)	103721 (58%)	21379 (67%)
No College	59790 (23%)	15958 (32%)	37344 (21%)	6488 (20%)
Unknown	55050 (21%)	12938 (26%)	38090 (21%)	4022 (13%)
Gender				
Female	105585 (41%)	23273 (47%)	70121 (39%)	12191 (38%)
Male	108522 (42%)	14854 (30%)	77850 (43%)	15818 (50%)
Other	1676 (0.6%)	580 (1.2%)	1017 (0.6%)	79 (0.2%)
Unknown	44456 (17%)	10488 (21%)	30167 (17%)	3801 (12%)

[†] Statistics presented: n (%)

Note: A total of 260,239 enrolled learners represented in the data for nine entry-level MOOCs, administered from 2017 to 2020 across 58 course-terms of students.

Additionally, 21% of students did not disclose their educational background. We chose to not drop these students, nor to use imputation methods to estimate educational background for this missing data. We included these students to conduct analysis on as large and reflective of the data set as possible, and because including the data as unknown did not detract from the analysis. Additionally, dropping incomplete cases biases data (Si & Reiter, 2013), and there is some evidence that students who are less likely to disclose demographic information may be more likely to come from underrepresented backgrounds (Jang & Vorderstrasse, 2019). We chose not to impute educational level data for several reasons. First, 21% is a substantial amount of data; while imputation methods have grown more sophisticated and can be quite accurate, they can still introduce bias and error (Si & Reiter, 2013; Lodder, 2014). Second, including the data labelled as Unknown is itself akin to constant imputation of a random categorical variable level. Third, the analysis most at risk of including the unknown variable label, the Gower distance–based clustering, is impacted in a predictable way. Namely, the Dice Coefficient component of Gower distance will separate clusters along binary dimensions of categorical variables. When Unknown is excluded from the analysis, four clusters are derived, two for each education background level, College Plus and No College; when Unknown is included, six clusters are derived, two for each education background level, College Plus, No College, and Unknown. This is further discussed in Appendix 1.

Table 2 presents the final data set of learners for which both gradebook and activity log data were present, which comprises the subset of 29,083 learners included in the cluster analyses. Limiting the analysis in this manner follows Wang et al. (2018),

which defines the "committed learner" as a student who submitted an assignment after the online analogue of the "add/drop period" of the course, representing a sample of 4.8% of total enrolled learners in their data set. In our sample, we broadened the "committed learner" definition to include any learner who submitted a graded assignment during a course sequential. The 29,083 learners included in cluster analyses represented 11.2% of the total data set. Limiting the sample to "committed learners" does carry some limitations, considered in the conclusion. Highly subsetted data is common across the MOOC literature and remains an area for improvement (Gardner & Brooks, 2018). Notably, the overall proportion of tertiary-educated learners decreased from enrolment to the committed learner sample, while non-tertiary educated learners increased. That said, college-educated learners still account for 52%.

Table 2: Descriptive Statistics for "Committed Learners" Included in Cluster Analyses

Characteristic	**Overall**, N = 29083	**USA**, N = 9708[1]	**Outside USA**, N = 15625[1]	**Unknown**, N = 3750[1]
Education_Level				
College Plus	15084 (52%)	3415 (35%)	9094 (58%)	2575 (69%)
No College	7951 (27%)	3778 (39%)	3463 (22%)	710 (19%)
Unknown	6048 (21%)	2515 (26%)	3068 (20%)	465 (12%)
Gender				
Female	13089 (45%)	4678 (48%)	6748 (43%)	1663 (44%)
Male	11093 (38%)	2952 (30%)	6436 (41%)	1705 (45%)
Other	241 (0.8%)	132 (1.4%)	97 (0.6%)	12 (0.3%)
Unknown	4660 (16%)	1946 (20%)	2344 (15%)	370 (9.9%)

[1] Statistics presented: n (%)

Note: A total of 29,083 learners with gradebook and activity data represented in nine entry-level MOOCs offered by the host university from 2017 to 2020, accounting for 58 course-terms of students.

3.2. Methods

Implementing cluster analysis methods involves careful consideration of data features to include, which distance measurements between those features to compute, as well as the selection of a clustering algorithm to group data into clusters based on those distance measurements (Kassambara, 2017; Kaufman & Rousseeuw, 1990). After experimenting with more than 30 different combinations of variables, distance measures, and algorithms, we settled on presenting two sets of exploratory cluster analyses to help answer the research questions.

There is no linear process to cluster analysis. Determining a distance measure to compute is dependent on the type of data available to cluster. Computing distances between observations across features is required to evaluate potential clusterability. If clusterability is determined, the selection of an algorithm depends on the distance measure computed, as well as whether that algorithm has corresponding software in a statistical learning package that enables it to derive clusters based on the distance measures selected. It is an interdependent process. There are a few commonly important steps, however, detailed by Kassambara (2017) and Kaufman and Rousseeuw (1990) and discussed below.

Selecting a Distance Measure: The first cluster analysis leverages Manhattan distances (Loohach & Garg, 2012), a method used commonly in the learning analytics literature (Kravvaris et al., 2016). Manhattan distance sums the absolute differences between observations across features, and requires interval data (Kassambara, 2017; Kaufman & Rousseeuw, 1990). Percent grade and a composite participation/performance metric are both interval observations between zero and one. Manhattan distance is also referred to as the L1 norm. Kizilcec et al. (2013) used Manhattan distances in their original cluster paper.

The second cluster analysis leverages Gower distance. Gower distance measures the similarity of observations across both numerical and categorical data (Ebbert & Dutke, 2020; Gower, 1971). Gower distance measures the Manhattan distances across interval data in a data set, and computes a Dice Coefficient across nominal data by first converting nominal variables of k categories into binary columns. Leveraging Gower distance allows for cluster analysis based a composite participation/performance metric, as well as educational background. The existing MOOC and learning analytic literature have rarely leveraged Gower distance to compute mixed-variable distances based on demographic and performance data to inform cluster analysis (Meaney, 2021).

Assessing Cluster Tendency: Clustering algorithms will find clusters in data arbitrarily if programmed to do so (Kassambara, 2017). Therefore, it is important to determine whether data is clusterable. One of the first ways of doing so is to take guidance from the existing literature, in which discovering clusters of learners is prominent (Li & Baker, 2018; Ferguson et al., 2015; Kizilcec et al., 2013). Second, several analytical methods exist to measure the relative dissimilarity across a data set. One of the most prominent is the Visual Assessment of Clustering Tendency, or VAT, which produces an ordered dissimilarity matrix (Kassambara, 2017). Ordered dissimilarity matrices are evaluated for potential indication of clusterability. These visualizations give some idea of how clusterable the data is by counting the number of dark-squared blocks along the diagonal axis (Kassambara, 2017), indicating distinct clusters. VAT is utilized for both cluster analysis explorations.

Determine Optimal Number of Clusters: Having determined that the data is clusterable, or potentially clusterable, it is important to then determine the number of clusters to explore. One common method is to evaluate the relative silhouette widths of various potential clusters to determine the best fit-value (Ferguson et al., 2015). Silhouette widths are an internal validation metric that measures how similar an observation is to its own cluster compared to its closest neighbouring cluster (Martin, 2016). A silhouette width close to 1 indicates that the object is well clustered. A silhouette width close to -1 indicates that the object is poorly clustered (Kassambara, 2017).

Another common method is the gap statistic method from Tibshirani et al. (2001). The gap statistic computes the total within intra-cluster variation for different numbers of clusters and compares this to the expected values of the total within intra-cluster variation of a null reference distribution of the data. The ideal number of clusters is estimated to be the value that maximizes the gap statistic; that is, the point at which the clustered structure is the furthest away from the random uniform distribution of points (Tibshirani et al., 2001).

Silhouette width analysis is utilized for both cluster analyses. The gap statistic method is not optimized for Gower distance, as it requires numeric variable inputs, and therefore is not utilized for the second set of analyses.

Implementing Clustering Algorithms: Selection of a clustering algorithm takes place alongside these steps, informed by the kind of data available, the distance measure computed, and the type of analysis pursued. Two of the most common types of clustering algorithms are partitioning and hierarchical methods (Kaufman & Rousseeuw, 1990). Partitioning methods construct k clusters by classifying data into k groups, where k is given by the researcher and usually determined separately. The groups contain none of the same data points. Hierarchical clustering divides or agglomerates data into groups as small as one to as large as the entire data set (Kaufman & Rousseeuw, 1990). In the learning analytics literature, partitioning methods are dominant (Khalil & Ebner, 2017; Arora et al., 2017; Kovanović et al., 2016; Ferguson et al., 2015; Ferguson & Clow, 2015; Kizilcec et al., 2013), though hierarchical methods have been used (Chen et al., 2015), as well as other techniques that form clusters (Anderson et al., 2014; Ramesh et al., 2014). Our analysis explores the more common approach of partitioning methods.

Within partitioning methods, two dominant approaches are k-means and k-medoids. Both methods work to minimize the within-cluster variation of objects, using two different approaches (Tibshirani, 2013). K-means is an algorithm that minimizes the sum of the squared error between data objects in a cluster and the centroid of that cluster. It begins by selecting random centre points of clusters and proceeds iteratively. K-medoids works in a similar fashion; however, instead of selecting an arbitrary centre point for the cluster, it selects an actual data point from the data set, and proceeds to minimize the sum of the dissimilarities between it and the observations assigned to its cluster. Each of these methods proceeds iteratively until the intra-cluster variation is minimized (Kassambara, 2017).

A k-medoids–based approach was selected for our cluster explorations for two reasons. First, we wanted to be consistent across both cluster explorations in terms of the methods used. K-means works well with Manhattan distances; however, it is not operable with the Gower distance. Second, it seemed sensible to base the clusters on central points representing actual observations in the data; in this case, an actual learner, as opposed to a mean point (Tibshirani, 2013). Therefore, algorithms based on partitioning around medoids were used, which can take either Manhattan distance or Gower distances. The primary k-medoids algorithm, PAM (Kaufman & Rousseeuw, 1990), is computationally expensive. Therefore, a more modern instantiation, CLARA, which leverages k-medoids–based clustering but does so on samples of the data set and is much faster, was used for the first cluster analysis with Manhattan distances. CLARA, however, is not optimized for Gower distance, so the traditional PAM algorithm was used for the second cluster analysis.

Post-Cluster Analysis: Once cluster analysis is conducted, the clusters are described, explored, and visualized. Tables and visualizations help illustrate the demographic distribution of education level and SES within clusters. These visualizations are paired with a univariate, multinomial logistic regression. Multinomial logistic regression models the log odds of nominal outcome variables, like the clusters, in relationship to the explanatory variables (Torres-Reyna, 2012; Long, 1997). This is not an exercise in predictive modelling, however, and there is no claim to causality. Indeed, the models only include one explanatory variable, education level or SES, and the Akaike Information Criterion (AIC) is quite high in absolute terms, though there is no comparison model to assess it relatively speaking. Instead, multinomial logistic regression is exploratory, and is utilized to make sense of the relative distributions of education levels across clusters. A similar analysis is conducted

with SES, on clusters formed from percent grade, the participation/performance metric, and education level. This will contribute to the literature's present ambiguous answer regarding whether demographic background variables influence student engagement and outcomes in MOOCs (Joksimović et al., 2018; Gardner & Brooks, 2018).

4. Analysis and Results

4.1. Analysis and Results: Manhattan Distance, CLARA-Based Clusters

In the first cluster analysis, learners are clustered based on Manhattan distances (Kravvaris et al., 2016; Loohach & Garg, 2012) and the CLARA algorithm (Schubert & Rousseeuw, 2019). The features of interest utilized to produce the clusters were percent grade and a composite participation/performance metric. After clustering, the relative distributions of education level across clusters are considered, which is then evaluated using multinomial logistic regression.

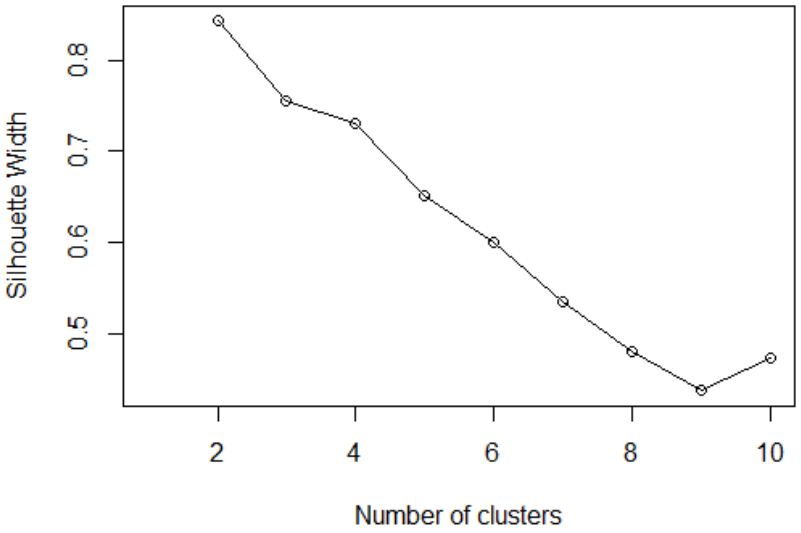

Figure 1: Silhouette plot of the CLARA clustering algorithm for k=2:10.

After features are selected for evaluation and Manhattan distances between the observations across the features are computed, clustering tendency is assessed (Kassambara, 2017). In addition to taking indications from the literature (Li & Baker, 2018), a Visual Assessment of Clustering Tendency (VAT) was produced, which presents an ordered dissimilarity matrix (Kassambara, 2017). Assessment of the VAT was somewhat ambiguous: there was no firm indication of distinct clusters. The pattern, however, was clearly non-random. Based on the existing literature, as well as the VAT, we inferred some grouping structure to the data. This was further assessed when determining the appropriate number of clusters to explore (Kassambara, 2017).

To determine the appropriate number of clusters to explore, average silhouette widths were computed. Similar to Ferguson and Clow (2015), silhouette width analysis was not particularly useful. Figure 1 shows the average silhouette widths observed when partitioning the data into a minimum of two groups and a maximum of ten. The silhouette width is greatest for two clusters, which may not be particularly meaningful, given that, typically, four subgroups of learners are often observed. A mild inflection point, which may indicate a salient number of clusters, can be seen at four clusters, and a sharper one at nine, though the silhouette width at nine clusters is below .5, indicating weak clustering. This is an ambiguous result. Therefore, another method for determining the best number of clusters was implemented: the gap statistic method (Tibshirani et al., 2001).

Figure 2 shows the output of implementing the gap statistic method with 100 Monte Carlo Bootstrapped random samples as the null reference set against a CLARA implementation of 30 samples of 1,000 data points from the MOOC data. The ideal number of clusters is determined to be three, with another levelling off of the gap statistic occurring at five; levelling can serve as an informal heuristic for potential investigation (Stack Exchange, 2014). While a rigorous approach, the result of three clusters differed slightly from the four to five clusters commonly found in the literature (Li & Baker, 2018).

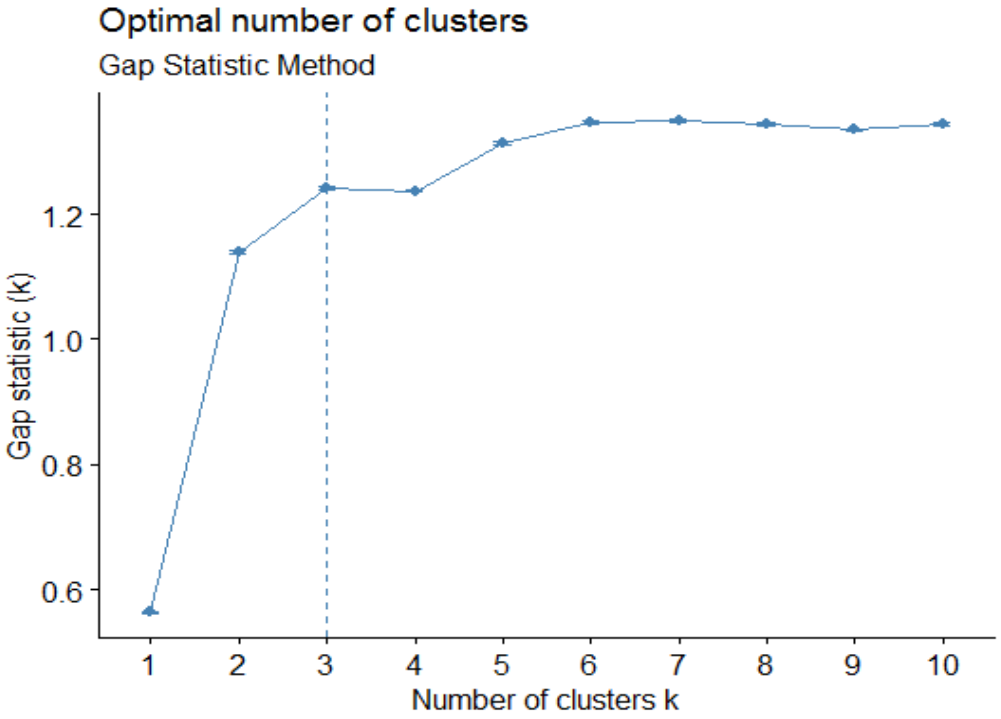

Figure 2: Estimates of the ideal number of CLARA-based clusters utilizing the gap statistic method.

4.1.1. Cluster Analysis and Descriptions

Given the ambiguity in the results, clusters three through five were explored. Four clusters yielded a result similar to clusters found in the literature of Disengagers, Auditors, Solvers, and All-rounders (Li & Baker, 2018). Five clusters broke the Disengagers into two groups with different engagement profiles. As a result, five clusters were selected for further analysis. Table 3 presents descriptive statistics of the five clusters. Leveraging common nomenclature from the literature (Li & Baker, 2018), the clusters are divided into the following categories: All-rounders, Auditors, Disengagers, Samplers, and Solvers. Figure 3 provides a visualization of the clusters along the percent grade axis and the participation/performance metric axis.

All-rounders accounted for 9.6% of the total sample. These learners achieved high marks in their courses, as reflected in their percent grade median of 89%, with an interquartile range of 81–94%. They actively engaged throughout the course, obtaining a participation/performance metric median score of 94%, with an interquartile range of 88–96%. They similarly had the highest median total event count of 4,586. The computed ratio of percent grade to participation/performance metric — called relative grade to engagement ratio — was .96, indicating consistent engagement. This is further contextualized below when compared to the Solver.

Auditors accounted for 7.1% of the total sample. These learners engaged with the course, obtaining a participation/performance median grade of 45%, with a median event count of 2,940. They scored less in their overall percent grade, with a median of 33%. This group did, however, have the largest range of interquartile median for both participation/performance and percent grade.

Disengagers accounted for 61% of the total sample. The largest group, these learners demonstrated little engagement and performance in the course, with a median percent grade of 0, and a participation/performance metric median of 2%.

Samplers accounted for 17% of the total sample. These learners dropped off early in the course, though they did register some engagement and achievement, with a median participation/performance grade of 16%, and a total percent grade of just 5% at the median. The event count median was 1,087, considerably higher than the Disengagers, who dropped off with far fewer events and had considerably less engagement and less achievement.

Solvers accounted for 4.9% of the total sample. These learners achieved high marks in their courses, though they did so while engaging considerably less. This is observed by their high percent grade median of 88%, and their relatively low participation/performance metric median of 46%. Their relative grade to engagement ratio was very high at 1.84, nearly double that of the All-rounders, meaning that these learners achieved roughly the same score while engaging with far fewer course sequentials, and likely in a more strategic, optimized way. They also had a lower event count median of 4,145.

Table 3: Descriptive Statistics for the Five Clusters of Learners: CLARA Algorithm

Characteristic	Overall, N = 29083	All-rounders, N = 2799[†]	Auditors, N = 2075[†]	Disengagers, N = 17817[†]	Samplers, N = 4971[†]	Solvers, N = 1421[†]
Education_Level						
College Plus	15084 (52%)	1514 (54%)	1012 (49%)	9452 (53%)	2550 (51%)	556 (39%)
No College	7951 (27%)	792 (28%)	692 (33%)	4379 (25%)	1505 (30%)	583 (41%)
Unknown	6048 (21%)	493 (18%)	371 (18%)	3986 (22%)	916 (18%)	282 (20%)
Part_and_Perf	0.05 (0.02, 0.23)	0.94 (0.88, 0.96)	0.45 (0.36, 0.59)	0.02 (0.02, 0.04)	0.16 (0.12, 0.22)	0.46 (0.38, 0.56)
Percent_Grade	0.01 (0.00, 0.11)	0.89 (0.81, 0.94)	0.33 (0.22, 0.44)	0.00 (0.00, 0.01)	0.05 (0.03, 0.09)	0.88 (0.79, 0.93)
Event_Count_total	547 (202, 1850)	4586 (3176, 6676)	2940 (1856, 4433)	261 (129, 534)	1087 (620, 1825)	4145 (3070, 5634)
Relative_Grade_to_Engagement_Ratio	0.12 (0.00, 0.61)	0.96 (0.89, 1.00)	0.68 (0.47, 0.91)	0.00 (0.00, 0.10)	0.37 (0.20, 0.51)	1.84 (1.58, 2.31)

[†] Statistics presented: n (%); median (IQR)

Note: All-rounders (9.6%), Auditors (7.1%), Disengagers (61%), Samplers (17%), and Solvers (4.9%). N = 29,083.

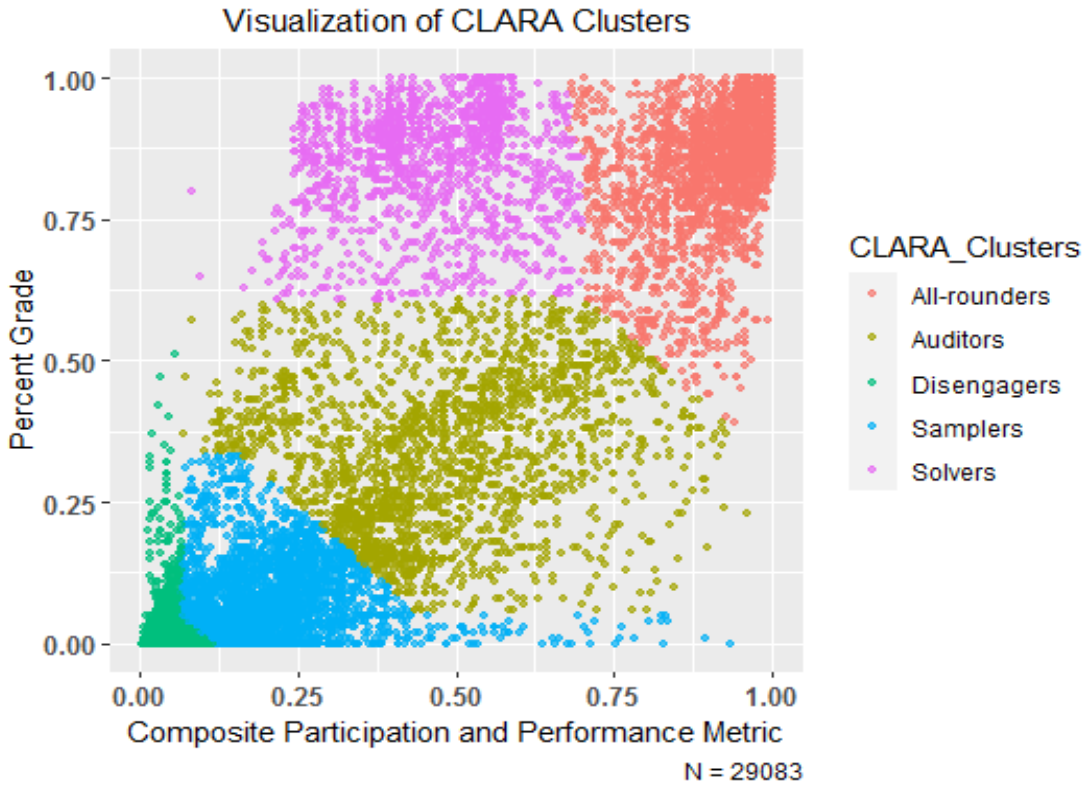

Figure 3: A representation of the five clusters determined by the CLARA algorithm. These include the commonly observed clusters in the literature: All-rounders, Auditors, Disengagers, Samplers, and Solvers. N = 29,083

4.1.2. Educational Backgrounds of Learners Across Clusters

When considering the educational backgrounds of the learners and how these subgroups dispersed across the clusters, more interesting insights emerged. Notably, College Plus learners are more likely to be Disengagers than No College learners, and No College learners are more likely to be Solvers than College Plus learners. At the same time, No College learners are more likely to be Samplers and Auditors than College Plus learners, which may have implications for learning design. This is considered further in the discussion. Table 4 presents the absolute and relative values of the different educational background levels distributed across the clusters. Figure 4 presents this visually.

Table 4: Distribution of Education Level Across the Five Clusters: CLARA Algorithm

Characteristic	**Overall**, N = 29083	**College Plus**, N = 15084[1]	**No College**, N = 7951[1]	**Unknown**, N = 6048[1]
CLARA_Clusters				
All-rounders	2799 (9.6%)	1514 (10%)	792 (10.0%)	493 (8.2%)
Auditors	2075 (7.1%)	1012 (6.7%)	692 (8.7%)	371 (6.1%)
Disengagers	17817 (61%)	9452 (63%)	4379 (55%)	3986 (66%)
Samplers	4971 (17%)	2550 (17%)	1505 (19%)	916 (15%)
Solvers	1421 (4.9%)	556 (3.7%)	583 (7.3%)	282 (4.7%)

[1] Statistics presented: n (%)

Note: N = 29,083.

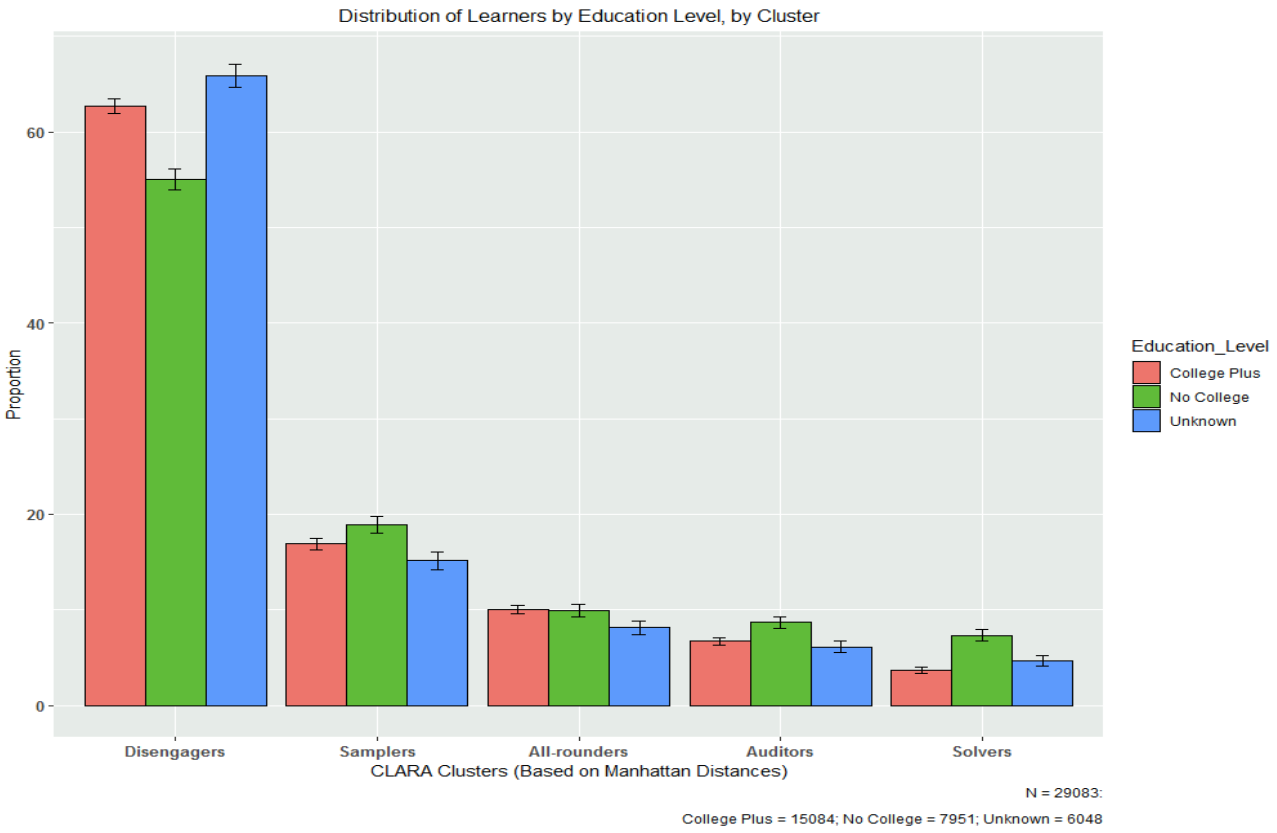

Figure 4: Distribution of educational background for students across five clusters determined by the CLARA algorithm. N = 29,083

Table 5 presents the relative risk ratios, displayed as the exponentiated value of the logit coefficients from a multinomial logistic regression (Torres-Reyna, 2012), where educational level is the explanatory variable, and cluster the outcome variable. The reference category is College Plus learners in the Disengagers cluster. The coefficients described are statistically significant at the .01 level. No College learners are 2.263 times as likely to be solvers, and 1.129 times as likely to be All-rounders, compared to the reference category. At the same time, they are 1.476 times as likely to be Auditors and 1.274 times as likely to be Samplers.

Table 5: Relative Risk Ratios: Education Level and Cluster

	Dependent variable:			
	All-rounders	Auditors	Samplers	Solvers
	(1)	(2)	(3)	(4)
Education_LevelNo College	1.129**	1.476***	1.274***	2.263***
	(0.048)	(0.053)	(0.037)	(0.062)
Education_LevelUnknown	0.772***	0.869**	0.852***	1.203**
	(0.055)	(0.064)	(0.043)	(0.076)
Constant	0.160***	0.107***	0.270***	0.059***
	(0.028)	(0.033)	(0.022)	(0.044)
Akaike Inf. Crit.	67,391.750	67,391.750	67,391.750	67,391.750
Note:				*p<0.1; **p<0.05; ***p<0.01

Note: Relative risk ratios, displayed as the exponentiated value of the logit coefficients from the multinomial logistic regression, where educational level is the explanatory variable, and cluster the outcome variable. College Plus, Disengagers are the reference group. N = 29,083.

4.2. Analysis and Results: Gower Distance, PAM-based Clusters

In the second cluster analysis, learners are clustered by the PAM algorithm based on Gower distances. The purpose of this investigation is twofold. First, it seeks to determine whether utilizing a categorical variable like education level is useful in the process of clustering. Second, it provides the opportunity to consider SES across clusters already controlling for education level, percent grade, and participation/performance. Additionally, few learning analytics papers on MOOCs have leveraged Gower distance measures to combine demographic and performance data to inform clusters, so the opportunity to implement it and possibly contribute a simple but meaningful mixed-type clustering use-case to the literature, especially useful for questions of fairness and equity, drove us to include these results.

Similar steps are followed as in the previous investigation. First, cluster tendency is evaluated, and then the ideal number of clusters is determined. Once the data is clustered, descriptive features of the clusters are presented. Then, SES is considered, for which only a small proportion of user data is available from the USA. SES is analyzed and presented in relation to the entire "committed learner" data set of 29,083 learners, despite only having SES data for 2,342 learners. When limiting the Gower distance–based cluster analysis to USA data only, six clusters are found, extremely similar to the six clusters found across the full sample of "committed learners"; thus, there was no need to differentiate the clustering results further. This is considered further in Appendix 1.

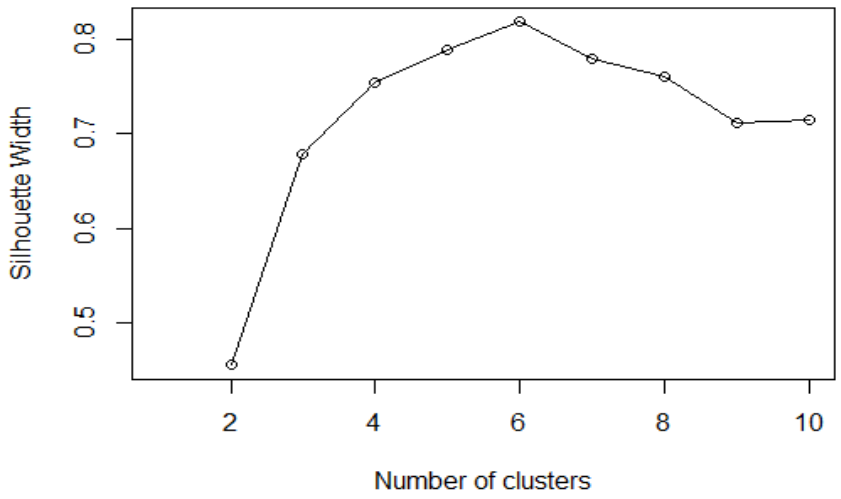

Figure 5: Silhouette plot of the PAM clustering algorithm for k=2:10. At six clusters, the average silhouette width for the data objects is above 0.8, indicating sound clusters

The initial impetus of this investigation was informed by the literature; specifically, common clusters of students are regularly identified in MOOC data (Li & Baker, 2018). In addition to the literature, the VAT technique was utilized. Again, the assessment of the VAT was ambiguous but clearly non-random, and the analysis proceeded to determine the appropriate number of clusters.

As opposed to the first cluster analysis in this paper, and in contrast to results in Ferguson and Clow (2015), silhouette analysis returned a highly interpretable and useful result. Figure 5 indicated a clear high point occurring at six clusters with an average silhouette width of above .8, indicating well-clustered observations. Based on this finding, the PAM clustering algorithm was implemented and specified to generate six clusters.

4.2.1. Cluster Analysis and Descriptions

Partitioning the data into six clusters yielded somewhat interesting yet not surprising results. Essentially, the algorithm splits the data into successful and unsuccessful students across educational backgrounds. This can, however, provide insight into how different educational groups are performing in comparison to each other, even when both are successful or not successful. Additionally, the common cluster types observed in the MOOC literature are not as evident in these clusters. While patterns similar to All-rounders and Disengagers are present across each education level, the Sampling, Auditing, and Solving cluster patterns are not observed. For these reasons, clusters are labelled in the following manner: College Plus, All-rounders; College Plus, Disengagers; No College, All-rounders; No College, Disengagers; and Unknown, All-rounders as well as Unknown, Disengagers. Table 6 presents descriptive statistics for the six clusters.

Table 6: Descriptive Statistics of the Six Clusters: PAM Algorithm

Characteristic	Overall, N = 29083	College Plus, All-rounders, N = 2411[1]	College Plus, Disengagers, N = 12673[1]	No College, All-rounders, N = 1710[1]	No College, Disengagers, N = 6241[1]	Unknown, All-rounders, N = 942[1]	Unknown, Disengagers, N = 5106[1]
Education_Level							
College Plus	15084 (52%)	2411 (100%)	12673 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
No College	7951 (27%)	0 (0%)	0 (0%)	1710 (100%)	6241 (100%)	0 (0%)	0 (0%)
Unknown	6048 (21%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	942 (100%)	5106 (100%)
Part_and_Perf	0.05 (0.02, 0.23)	0.87 (0.60, 0.95)	0.04 (0.02, 0.09)	0.71 (0.52, 0.94)	0.04 (0.02, 0.11)	0.78 (0.54, 0.92)	0.03 (0.02, 0.08)
Percent_Grade	0.01 (0.00, 0.11)	0.87 (0.74, 0.93)	0.00 (0.00, 0.03)	0.84 (0.64, 0.92)	0.00 (0.00, 0.04)	0.85 (0.66, 0.92)	0.00 (0.00, 0.01)
Event_Count_total	547 (202, 1850)	4122 (2864, 6084)	381 (171, 881)	4384 (3133, 6191)	455 (179, 1084)	4444 (3162, 6224)	366 (141, 849)
Relative_Grade_to_Engagement_Ratio	0.12 (0.00, 0.61)	0.98 (0.89, 1.13)	0.00 (0.00, 0.35)	0.99 (0.86, 1.61)	0.00 (0.00, 0.38)	0.99 (0.86, 1.48)	0.00 (0.00, 0.23)

[1] Statistics presented: n (%); median (IQR)

Note: College Plus All-rounders (8.3%), College Plus Disengagers (44%), No College All-rounders (5.9%), No College Disengagers (21%), Unknown All-rounders (3.2%), and Unknown Disengagers (18%). Total N = 29,083.

College Plus, All-rounders accounted for 8.3% of the total sample. This cluster is composed of students with a tertiary degree who successfully completed their courses with a high degree of participation across course sequentials. They had a median percent grade of 87%, and a median participation/performance metric of 87%.

College Plus, Disengagers accounted for 44% of the total sample, the largest group among all the clusters. These learners had a median percent grade of zero, a median participation/performance metric of 4%, and average event counts of 381. This group appears very similar to the Disengagers from the first set of clusters.

No College, All-rounders accounted for 5.9% of the sample. They behaved similarly to College Plus All-rounders, with a median percent grade of 84%, and a median participation/performance metric of 71%. While still successful in their courses, these learners had lower overall percent grades and participation/performance metrics than the College Plus All-rounders. At the same time, they had a greater median event count than the College Plus All-rounders.

No College, Disengagers accounted for 21% of the total sample. Like the College Plus Disengagers, these learners obtained a median percent grade of zero and a median participation/performance metric of 4%. They averaged 451 events, and fit patterns similar to the Disengagers from the first cluster analysis.

Unknown, All-rounders accounted for 3.2% of the sample. This group obtained a median percent grade of 84%, a median participation/performance metric of 78%, and a median event count of 4,444. These numbers place learners in this cluster in between the College Plus and No College All-rounders. At the median, these learners achieved and participated slightly less than the College Plus All-rounder cohort, and slightly more than the No College All-rounder cohort.

Unknown, Disengagers accounted for 18% of the sample. With very similar patterns to the first two groups of Disengagers from this cluster analysis, and similar to the Disengagers from the first cluster analysis, these learners obtained a median percent grade of zero, a median participation/performance metric of 3%, and 366 events at the median.

4.2.2. Socioeconomic Backgrounds of Learners across Clusters

While data for only a small sample of survey completers was used to approximate SES status, enriching the clusters with this data provides deeper insight than just considering education level itself. It allows us to examine the SES distribution across the clusters, controlling for educational status, percent grade, and the participation/performance metric. Table 7 presents the distribution of SES status across the clusters, where that data is available. Similarly, Figure 6 provides a visualization of the distribution of SES across clusters. Parameter estimates for mid-high- and low-SES have considerably wide confidence intervals resultant from the relatively small sample size. As discussed in Appendix 1, limiting the initial Gower distance–based cluster analysis to USA-only data yielded extremely similar results to the Gower distance–based cluster analysis on the full data set; therefore, SES is presented in relation to the full data set. That such a limited amount of USA-only SES data is available is a limitation of the study and the findings are presented as exploratory. Nevertheless, we included this analysis because the learning analytics literature to date provides little insight into the relationship between SES and cluster sorting among MOOC users. Examining SES while controlling for behaviour type and educational background presents a novel research method to consider that may be relevant for researchers interested in expanding coverage in learning analytics of questions dealing with fairness and equity.

Table 7: Distribution of SES Across the Six Clusters: PAM Algorithm

Characteristic	**Overall**, N = 29083	**Unknown SES**, N = 26741[1]	**Low SES**, N = 458[1]	**Mid-High SES**, N = 1884[1]
PAM_Clusters				
College Plus, All-rounders	2411 (8.3%)	2284 (8.5%)	21 (4.6%)	106 (5.6%)
College Plus, Disengagers	12673 (44%)	12188 (46%)	74 (16%)	411 (22%)
No College, All-rounders	1710 (5.9%)	1364 (5.1%)	69 (15%)	277 (15%)
No College, Disengagers	6241 (21%)	5640 (21%)	120 (26%)	481 (26%)
Unknown, All-rounders	942 (3.2%)	685 (2.6%)	58 (13%)	199 (11%)
Unknown, Disengagers	5106 (18%)	4580 (17%)	116 (25%)	410 (22%)
[1] Statistics presented: n (%)				

Note: Total N = 29,083.

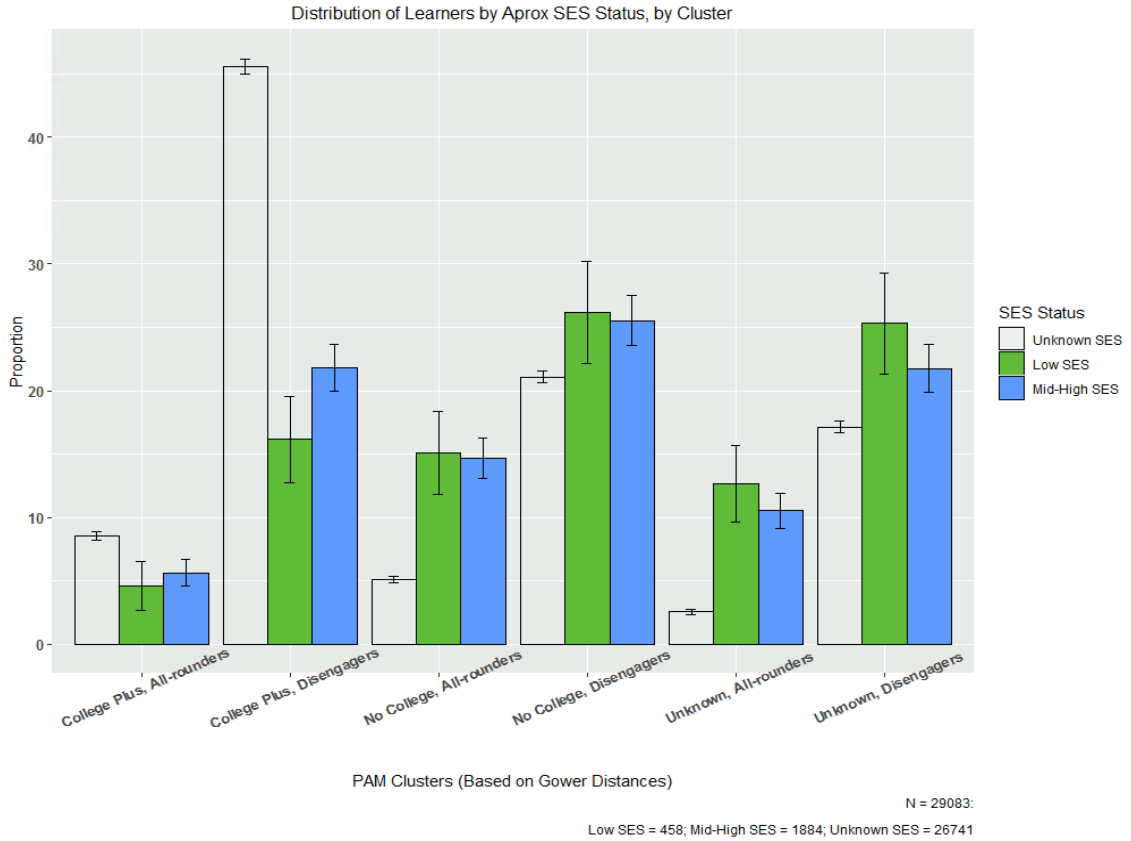

Figure 6: The distribution of known SES across clusters determined in part by educational level. Total N = 29,083

Table 8 presents the relative risk ratios, shown as the exponentiated value of the logit coefficients from the multinomial logistic regression (Torres-Reyna, 2012), where SES is the explanatory variable, and cluster is the outcome variable. The reference category is College Plus All-rounders from mid-high-SES backgrounds. Low-SES learners appear to be just as likely as their mid-high-SES counterparts to be in any of the clusters, as none of the relative risk ratios registers statistical significance. This could be a sample size limitation. When looking at the coefficients, however, it appears at least possible that while a larger sample could validate low-SES learners as more likely to be in the No College or Unknown education level clusters, they appear to be more or less evenly distributed across both the successful All-rounder group and the unsuccessful Disengager groups.

Table 8: Relative Risk Ratios: SES and Cluster

	Dependent variable:				
	College Plus, Disengagers	No College, All-rounders	No College, Disengagers	Unknown, All-rounders	Unknown, Disengagers
	(1)	(2)	(3)	(4)	(5)
bi_SES_status_finalUnknown SES	1.376***	0.229***	0.544***	0.160***	0.518***
	(0.111)	(0.119)	(0.110)	(0.128)	(0.112)
bi_SES_status_finalLow SES	0.909	1.257	1.259	1.471	1.428
	(0.270)	(0.274)	(0.260)	(0.282)	(0.261)
Constant	3.878***	2.613***	4.538***	1.877***	3.868***
	(0.109)	(0.114)	(0.107)	(0.120)	(0.109)
Akaike Inf. Crit.	85,200.670	85,200.670	85,200.670	85,200.670	85,200.670
Note:					*p<0.1; **p<0.05; ***p<0.01

Note: The relative risk ratios, shown as the exponentiated value of the logit coefficients from the multinomial logistic regression, where SES is the explanatory variable, and cluster is the outcome variable. College Plus All-rounders from mid-high-SES backgrounds are the reference group. N = 29083.

5. Discussion, Conclusions, and Limitations

This paper aimed to explore how learners from traditionally underrepresented backgrounds among MOOC users — students without a college degree, and students from lower socioeconomic backgrounds — are utilizing MOOCs intentionally designed to be more inclusive by a research-intensive university in the USA. It does so by utilizing a common learning analytic technique, cluster analysis, and exploring how clusters are differentiated by demographic subgroups. Additionally, a novel computational technique, Gower distance, was used to explore mixed-variable clustering so that performance and demographic data could both be used to form clusters.

Results indicated that learners in the entry-level MOOCs analyzed sorted into behaviour patterns prototypical in the MOOC literature: All-rounders, Auditors, Disengagers, Samplers, and Solvers.

More interesting, however, was how users from different demographic backgrounds sorted into these clusters. The first cluster analysis indicated that, in these specific entry-level MOOCs, traditionally underrepresented learners on the dimension of education level are more likely to sort into the commonly observed successful subgroups of All-rounders and Solvers compared to their peers who were better educated Disengagers. At the same time, underrepresented learners on the dimension of education level are also more likely to sort into the Auditing and Sampling clusters, potentially indicating a need for timely, targeted support. Further analysis seeking to understand the motivations of all learners is needed to make better sense of these outcomes. Greater understanding of the top of the MOOC entry funnel is also needed. Even among the entry-level MOOCs analyzed in this paper, the top of the MOOC funnel was still disproportionately populated by already well-educated learners. Importantly, this pattern is not just an enrolment issue either since it persisted even in the subset of committed learners who submitted at least one assignment.

The second cluster analysis pursued inquiries into whether demographic variables had a relationship to performance and achievement in the MOOCs by constructing clusters based on percent grade, computed participation/performance, and education level. These clusters were sorted into successful and unsuccessful learners across the education levels present — College Plus, No College, and Unknown — for a total of six clusters. When these clusters were enriched with approximated SES status data for a small subset of learners in the USA, multinomial logistic regression indicated that low-SES learners were just as likely as their mid-high-SES peers to sort into any of the other clusters. These insights are asserted with caution, as sample size issues limited the breadth and depth of analysis capable along the SES dimension.

Furthermore, concerning our methods, results indicated that utilizing a categorical variable like education level as a feature to derive clusters and a distance measure like Gower (Ebbert & Dutke, 2020; Gower, 1971) may be a worthwhile approach to consider. Specifically, utilizing mixed-variable features to cluster data may reveal subgroups of underrepresented learners potentially more amenable to support and remediation interventions; for example, the group of No College Disengagers observed in the second cluster analysis presented.

Regarding the existing literature, these results have several implications. First, while there is no consensus regarding the relationship between demographic variables like education level and MOOC engagement, some literature does indicate that traditionally underrepresented learners can be just as successful as their better-represented, higher-educated peers, especially in courses more intentionally designed for their demographic (Lambert, 2020; Wang et al., 2018; Goldberg et al., 2015). The results in this paper offer provisional support for these conclusions and offer some evidence that observed patterns of better-educated learners performing better in MOOCs (Kizilcec & Halawa, 2015; Engle et al., 2015) do not have to be the case. Similarly, while the analysis of SES data was limited by sample size, the observed evidence of learners from high-SES backgrounds performing better in MOOCs (Ganelin & Chuang, 2019; Hansen & Reich, 2015) should not be assumed either. More research is needed to determine what specifically enabled these outcomes. Importantly, however, these outcomes suggest that the potential MOOC "pivot" to focus on providing continuing education to already well-educated professionals may be premature (Reich & Ruipérez-Valiente, 2019). Additionally, results in this paper suggest that taking a more explicit approach to investigating and analyzing MOOC data across demographic variables, utilizing a mixed-variable distance measure like Gower, would be a worthwhile and potentially promising way to move MOOCs back toward their original mission and to support learning analytics more generally to expand coverage on questions of fairness and equity.

5.1. Limitations

There are considerable limitations to this study worth noting that qualify the conclusions. First, using limited activity-based features constrained the analysis. Categorizing total event counts into further specified activities like video-watching or peer-to-peer engagement could have made the clusters and the engagement descriptions richer. Second, the analysis could have been improved by considering whether the specific courses were associated with differential achievement and engagement, as indicated elsewhere in the literature (Ferguson & Clow, 2015). Third, limiting the cluster analyzed sample to committed learners further narrowed the scope of the results and the claims that can be made, a common issue in learning analytic research (Gardner & Brooks, 2018). Conducting multinomial logistic regression with only one explanatory variable significantly limits the scope of what the analysis can claim. The bias resultant from selection into completing both enrolment demographic

questionnaires and optional survey data represents another limitation to the analysis. Underrepresented populations, particularly along racial lines, have been found to be less likely to consent to engage in studies, as well as less likely to complete surveys, in web-based research. This means that the results could further reflect and embed those biases into the conclusions (Jang & Vorderstrasse, 2019). Furthermore, it is possible that more engaged learners were more likely to complete these questionnaires and surveys (Kizilcec & Schneider, 2015). Given that educational level was a variable used to cluster the data with the Gower distance metric, the inclusion of unknown education level data renders two of the six Gower distance–based clusters defined by the property of their educational level being unknown. More sophisticated imputation methods could have been pursued. That said, including the unknown education level data did not detract from analysis and led to a predictable outcome, discussed further in Appendix 1.

Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

Funding for this project was provided, in part, by the Gates Cambridge Trust. Resources and support were provided, in part, by the University of Cambridge and Arizona State University.

Acknowledgments

Special thanks to Ricardo Sabates of the University of Cambridge. As the corresponding author's doctoral adviser, Ricardo played an instrumental role in developing the thinking and methodologies that underpin this paper. Thanks to Rupert Wegerif of the University of Cambridge and Rebecca Ferguson of the Open University, thesis committee members of the corresponding author who provided feedback to strengthen this paper. Additional thanks to Mohammed Khalil of the University of Bergen, and the editors and reviewers of the special issue on Fairness, Equity, and Responsibility of the Journal of Learning Analytics for their constructive comments and support. An earlier draft of this paper was published as short paper at the ACM Conference on Learning @ Scale in 2022, and the authors are also grateful to the reviewers and program chairs of that conference for the helpful feedback.

References

Anderson, A., Huttenlocher, D., Kleinberg, J., & Leskovec, J. (2014). Engaging with massive online courses. *Proceedings of the 23rd International Conference on World Wide Web* (WWW '14), 7–11 April 2014, Seoul, Republic of Korea (pp. 687–698). ACM Press. https://doi.org/10.1145/2566486.2568042

Arora, S., Goel, M., Sabitha, A. S., & Mehrotra, D. (2017). Learner groups in massive open online courses. *American Journal of Distance Education*, *31*(2), 80–97. https://doi.org/10.1080/08923647.2017.1300461

Brooks, C., Thompson, C., & Teasley, S. (2015). Who you are or what you do: Comparing the predictive power of demographics vs. activity patterns in massive open online courses (MOOCs). *Proceedings of the 2nd ACM Conference on Learning @ Scale* (L@S 2015), 14–18 March 2015, Vancouver, BC, Canada (pp. 245–248). ACM Press. https://doi.org/10.1145/2724660.2728668

Carnegie Classifications of Institutions of Higher Education. (2017). Basic classification description. https://carnegieclassifications.iu.edu/classification_descriptions/basic.php

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, *61*(6), 1–36. https://doi.org/10.18637/jss.v061.i06

Chen, B., Håklev, S., Harrison, L., Najafi, H., & Rolheiser, C. (2015). How do MOOC learners' intentions relate to their behaviors and overall outcomes? *Proceedings of the American Educational Research Association Annual Conference* (AERA 2015), 16–20 April 2015, Chicago, IL, USA. https://www.researchgate.net/publication/316605290_How_do_MOOC_learners'_intentions_relate_to_their_behaviors_and_overall_outcomes#fullTextFileContent

Chetty, R., Friedman, J. N., Saez, E., Turner, N., & Yagan, D. (2017). Mobility report cards: The role of colleges in intergenerational mobility (No. w23618). National Bureau of Economic Research. https://doi.org/10.3386/w23618

Deng, R., Benckendorff, P., & Gannaway, D. (2017). Understanding learning and teaching in MOOCs from the perspectives of students and instructors: A review of literature from 2014 to 2016. *Proceedings of the 5th European MOOCs Stakeholders Summit* (EMOOCs 2017), 22–26 May 2017, Madrid, Spain (pp. 176–181). Springer. https://doi.org/10.1007/978-3-319-59044-8_20

Deng, R., Benckendorff, P., & Gannaway, D. (2019). Progress and new directions for teaching and learning in MOOCs. *Computers & Education*, *129*, 48–60. https://doi.org/10.1016/j.compedu.2018.10.019

Dillahunt, T. R., Wang, B. Z., & Teasley, S. (2014). Democratizing higher education: Exploring MOOC use among those who cannot afford a formal education. *The International Review of Research in Open and Distributed Learning*, *15*(5), 177–196. https://doi.org/10.19173/irrodl.v15i5.1841

Dowle, M., Srinivasan, A., Gorecki, J., Chirico, M., Stetsenko, P., Short, T., Lianoglou, S., Antoyan, E., Bonsch, M., Parsonage, H., Ritchie, S., Ren, K., Tan, X., Saporta, R., Seiskari, O., Dong, X., Lang, M., Iwasaki, W., Wenchel, S., ... Schwen, B. (2022). Extension of 'data.frame.' https://cran.r-project.org/web/packages/data.table/data.table.pdf

Ebbert, D., & Dutke, S. (2020). Patterns in students' usage of lecture recordings: A cluster analysis of self-report data. *Research in Learning Technology*, *28*. https://doi.org/10.25304/rlt.v28.2258

Engle, D., Mankoff, C., & Carbrey, J. (2015). Coursera's introductory human physiology course: Factors that characterize successful completion of a MOOC. *The International Review of Research in Open and Distributed Learning*, *16*(2), 46–68. https://doi.org/10.19173/irrodl.v16i2.2010

Escobari, M., Seyal, I., & Meaney, M. J. (2019*).* Realism about reskilling: Upgrading the career prospects of America's low-wage workers. The Brookings Institution. https://www.brookings.edu/research/realism-about-reskilling/

Ferguson, R., & Clow, D. (2015). Examining engagement: Analysing learner subpopulations in massive open online courses (MOOCs). *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (LAK '15), 16–20 March 2015, Poughkeepsie, NY, USA (pp. 51–58). ACM Press. https://doi.org/10.1145/2723576.2723606

Ferguson, R., Clow, D., Beale, R., Cooper, A. J., Morris, N., Bayne, S., & Woodgate, A. (2015). Moving through MOOCS: Pedagogy, learning design and patterns of engagement. *Proceedings of the 10th European Conference on Technology Enhanced Learning: Design for Teaching and Learning in a Networked World* (EC-TEL 2015), 15–17 September 2015, Toledo, Spain (pp. 70–84). Lecture Notes in Computer Science, vol. 9307. Springer. https://doi.org/10.1007/978-3-319-24258-3_6

Ganelin, D., & Chuang, I. (2019). IP geolocation underestimates regressive economic patterns in MOOC usage. *Proceedings of the 2019 11th International Conference on Education Technology and Computers* (ICETC 2019), 28–31 October 2019, Amsterdam, Netherlands (pp. 268–272). ACM Press. https://doi.org/10.1145/3369255.3369301

Gardner, J., & Brooks, C. (2018). Student success prediction in MOOCs. *User Modeling and User-Adapted Interaction*, *28*, 127–203. https://doi.org/10.1007/s11257-018-9203-z

Goldberg, L. R., Bell, E., King, C., O'Mara, C., McInerney, F., Robinson, A., & Vickers, J. (2015). Relationship between participants' level of education and engagement in their completion of the Understanding Dementia massive open online course. *BMC medical education*, *15*, 60. https://doi.org/10.1186/s12909-015-0344-z

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, *27*(4), 857–871. https://doi.org/10.2307/2528823

Greene, J. A., Oswald, C. A., & Pomerantz, J. (2015). Predictors of retention and achievement in a massive open online course. *American Educational Research Journal*, *52*(5), 925–955. https://doi.org/10.3102/0002831215584621

Guo, P. J., & Reinecke, K. (2014). Demographic differences in how students navigate through MOOCs. *Proceedings of the 1st ACM Conference on Learning @ Scale* (L@S 2014), 4–5 March 2014, Atlanta, GA, USA (pp. 21–30). ACM Press. https://doi.org/10.1145/2556325.2566247

Hansen, J. D., & Reich, J. (2015). Democratizing education? Examining access and usage patterns in massive open online courses. *Science*, *350*(6265), 1245–1248. https://doi.org/10.1126/science.aab3782

Jang, M., & Vorderstrasse, A. (2019). Socioeconomic status and racial or ethnic differences in participation: Web-based survey. *JMIR Research Protocols*, *8*(4), e11865. https://www.researchprotocols.org/2019/4/e11865/

Joksimović, S., Kovanović, V., & Dawson, S. (2019). The journey of learning analytics. *HERDSA Review of Higher Education*, *6*, 37–63. https://www.herdsa.org.au/herdsa-review-higher-education-vol-6/37-63

Joksimović, S., Poquet, O., Kovanović, V., Dowell, N., Mills, C., Gašević, D., Dawson, S., Graesser, A. C., & Brooks, C. (2018). How do we model learning at scale? A systematic review of research on MOOCs. *Review of Educational Research*, *88*(1), 43–86. https://doi.org/10.3102/0034654317740335

Jordan, K. (2014). Initial trends in enrolment and completion of massive open online courses. *The International Review of Research in Open and Distributed Learning*, *15*(1), 133–160. https://doi.org/10.19173/irrodl.v15i1.1651

Kassambara, A. (2017). *Practical guide to cluster analysis in R: Unsupervised machine learning*. CreateSpace Independent Publishing Platform.

Kassambara, A., & Mundt, F. (2016). Package 'factoextra': Extract and visualize the results of multivariate data analyses. https://cran.microsoft.com/snapshot/2016-11-30/web/packages/factoextra/factoextra.pdf

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. John Wiley & Sons. https://onlinelibrary.wiley.com/doi/book/10.1002/9780470316801

Khalil, M., & Ebner, M. (2017). Clustering patterns of engagement in massive open online courses (MOOCs): The use of learning analytics to reveal student categories. *Journal of Computing in Higher Education, 29*(1), 114–132. https://doi.org/10.1007/s12528-016-9126-9

Kizilcec, R. F., & Halawa, S. (2015). Attrition and achievement gaps in online learning. *Proceedings of the 2nd ACM Conference on Learning @ Scale* (L@S 2015)*,* 14–18 March 2015, Vancouver, BC, Canada (pp. 57–66). ACM Press. https://doi.org/10.1145/2724660.2724680

Kizilcec, R. F., Piech, C., & Schneider, E. (2013). Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge* (LAK '13), 8–12 April 2013, Leuven, Belgium (pp. 170–179). ACM Press. https://doi.org/10.1145/2460296.2460330

Kizilcec, R. F., & Schneider, E. (2015). Motivation as a lens to understand online learners: Toward data-driven design with the OLEI scale. *ACM Transactions on Computer–Human Interaction*, *22*(2), 1–24. https://doi.org/10.1145/2699735

Kovanović, V., Joksimović, S., Gašević, D., Owers, J., Scott, A.-M., & Woodgate, A. (2016). Profiling MOOC course returners: How does student behaviour change between two course enrollments? *Proceedings of the 3rd ACM Conference on Learning @ Scale* (L@S 2016)*,* 25–28 April 2016, Edinburgh, Scotland (pp. 269–272). ACM Press. https://doi.org/10.1145/2876034.2893431

Kravvaris, D., Kermanidis, K. L., & Ntanis, G. (2016). How MOOCs link with social media. *Journal of the Knowledge Economy*, *7*, 461–487. https://doi.org/10.1007/s13132-014-0219-2

Lambert, S. R. (2020). Do MOOCs contribute to student equity and social inclusion? A systematic review 2014–18. *Computers & Education*, *145*, 103693. https://doi.org/10.1016/j.compedu.2019.103693

Li, Q., & Baker, R. (2018). The different relationships between engagement and outcomes across participant subgroups in massive open online courses. *Computers & Education*, *127*, 41–65. https://doi.org/10.1016/j.compedu.2018.08.005

Lodder, P. (2014). To impute or not impute: That's the question. In G. J. Mellenbergh & H. J. Adèr (Eds.), *Advising on research methods: Selected topics 2013*. Johannes van Kessel Publishing. https://www.paultwin.com/wp-content/uploads/Lodder_1140873_Paper_Imputation.pdf

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Sage Publications. https://us.sagepub.com/en-us/nam/regression-models-for-categorical-and-limited-dependent-variables/book6071

Loohach, R., & Garg, K. (2012). Effect of distance functions on k-means clustering algorithm. *International Journal of Computer Applications*, *49*(6), 7–9. https://doi.org/10.5120/7629-0698

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., Studer, M., Roudier, P., Gonzalez, J., Kozlowski, K., Schubert, E., & Murphy, K. (2022). "Finding groups in data": Cluster analysis extended. http://mirrors.ucr.ac.cr/CRAN/web/packages/cluster/cluster.pdf

Martin, D. (2016). Clustering mixed data types in R. https://dpmartin42.github.io/posts/r/cluster-mixed-types

Meaney, M. (2021). *Essays on the design of inclusive learning in Massive Open Online Courses, and implications for educational futures* [Unpublished doctoral dissertation]. University of Cambridge. https://doi.org/10.17863/CAM.76128

Meaney, M. J., & Fikes, T. (2019). Early-adopter iteration bias and research-praxis bias in virtual learning environments. *Companion Proceedings of the 9th International Conference on Learning Analytics and Knowledge* (LAK '19), 4–8 March 2019, Tempe, Arizona (pp. 513–519). Society for Learning Analytics Research (SoLAR). https://www.solaresearch.org/wp-content/uploads/2019/08/LAK19_Companion_Proceedings.pdf

Meaney, M., & Fikes, T. (2022). Adding a demographic lens to cluster analysis of participants in entry-level massive open online courses (MOOCs). *Proceedings of the 9th ACM Conference on Learning @ Scale* (L@S 2022)*,* 1–3 June 2022, New York, NY, USA (pp. 355–359). ACM Press. https://doi.org/10.1145/3491140.3528306

R Core Team. (2019). *R: A language and environment for statistical computing*. The R Foundation for Statistical Computing. https://www.R-project.org/

Ramesh, A., Goldwasser, D., Huang, B., Daume III, H., & Getoor, L. (2014). Learning latent engagement patterns of students in online courses. *Proceedings of the 28th Conference on Artificial Intelligence* (AAAI-18), 27–31 July 2014, Québec City, Québec, Canada (pp. 1272–1278). AAAI Press. https://doi.org/10.1609/aaai.v28i1.8920

Reich, J., & Ruipérez-Valiente, J. A. (2019). The MOOC pivot. *Science*, *363*(6423), 130–131. https://doi.org/10.1126/science.aav7958

Rohs, M., & Ganz, M. (2015). MOOCs and the claim of education for all: A disillusion by empirical data. *The International Review of Research in Open and Distributed Learning*, *16*(6), 1–19. https://doi.org/10.19173/irrodl.v16i6.2033

Savje, F. (2021). Distances. GitHub, Inc. https://github.com/fsavje/distances

Schubert, E., & Rousseeuw, P. J. (2019). Faster *k*-medoids clustering: Improving the PAM, CLARA, and CLARANS algorithms. *Proceedings of the 12th International Conference on Similarity Search and Applications* (SISAP 2019), 2–4 October 2019, Newark, NJ, USA (pp. 171–187). Springer. https://doi.org/10.1007/978-3-030-32047-8_16

Si, Y., & Reiter, J. P. (2013). Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics, 38*(5), 499–521. https://doi.org/10.3102/1076998613480394

Stack Exchange. (2014, April 26). How should I interpret GAP statistic? [Online forum post]. https://stats.stackexchange.com/questions/95290/how-should-i-interpret-gap-statistic/210704

Stich, A. E., & Reeves, T. D. (2017). Massive open online courses and underserved students in the USA. *The Internet and Higher Education*, *32*, 58–71. https://doi.org/10.1016/j.iheduc.2016.09.001

Tibshirani, R. (2013). Clustering 1: *K*-means, *K*-medoids. https://www.stat.cmu.edu/~ryantibs/datamining/lectures/04-clus1.pdf

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411–423. https://doi.org/10.1111/1467-9868.00293

Torres-Reyna, O. (2012). Getting started in Logit and ordered Logit regression. Princeton University. http://dss.princeton.edu/training/Logit.pdf

Wang, Y., Fikes, T. G., & Pettyjohn, P. (2018). Open scale courses: Exploring access and opportunity for less-educated learners. *Proceedings of the 2018 Learning with MOOCS Conference* (LWMOOCS '18), 26–28 September 2018, Madrid, Spain (pp. 102–105). IEEE. https://doi.org/10.1109/LWMOOCS.2018.8534667

Wickham, H. (2022). Easily install and load the 'tidyverse' [R package tidyverse version 1.3.2]. Comprehensive R Archive Network (CRAN). https://cran.r-project.org/web/packages/tidyverse/index.html

Zamecnik, A., Kovanović, V., Joksimović, S., & Liu, L. (2022). Exploring non-traditional learner motivations and characteristics in online learning: A learner profile study. *Computers and Education: Artificial Intelligence*, *3*, 100051. https://doi.org/10.1016/j.caeai.2022.100051

Zhang, Q., Bonafini, F. C., Lockee, B. B., Jablokow, K. W., & Hu, X. (2019). Exploring demographics and students' motivation as predictors of completion of a massive open online course. *The International Review of Research in Open and Distributed Learning*, *20*(2), 140–161. https://doi.org/10.19173/irrodl.v20i2.3730

Zhenghao, C., Alcorn, B., Christensen, G., Eriksson, N., Koller, D., & Emanuel, E. J. (2015, September 22). Who's benefiting from MOOCs, and why. *Harvard Business Review*. https://hbr.org/2015/09/whos-benefiting-from-moocs-and-why

APPENDICES

Appendix 1: Supplementary Analysis

Of the 29,083 learners included in the cluster analyses presented in this paper, some 21% did not provide educational background data. The Gower distance–based cluster analyses utilize education background as a key variable in computing the distance between clusters. Initially, we did not consider imputing the data because we did not think that leaving the data as unknown detracted from the analysis. Clearly educational background levels did indeed impact the clustering, as represented by the other clusters. Determining whether this was the case was the primary objective of the analysis.

The analysis most at risk of including the unknown variable level, the Gower distance–based clustering, is impacted in a predictable way. Namely, the dice coefficient component of Gower distance will separate clusters along binary dimensions of categorical variables. This is indicated when deriving the number of ideal clusters to utilize in the analysis via evaluating the silhouette widths. When unknown is excluded from the analysis, four clusters are derived, two for each education background level, College Plus and No College. This is represented in Figure A.1 and Table A.1, which confirm that the four clusters contain groups of higher achieving and lower-achieving College Plus learners, and groups of higher achieving and lower achieving No College learners.

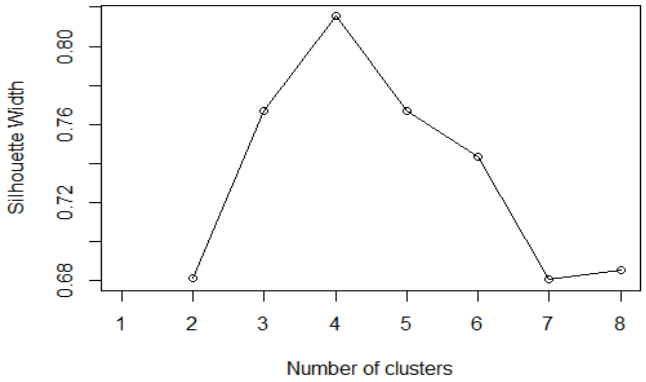

Figure A.1: Silhouette plot of the PAM clustering algorithm for k=2:8, excluding unknown education level data. At four clusters, the average silhouette width for the data objects is above 0.8, indicating sound clusters. Total N = 23,305, excluding all unknown education level data

Table A.1: Descriptive Statistics of the Four Clusters: PAM Algorithm

Characteristic	**Overall**, N = 23035	**College Plus, All-rounders**, N = 2411[1]	**College Plus, Disengagers**, N = 12673[1]	**No College, All-rounders**, N = 1710[1]	**No College, Disengagers**, N = 6241[1]
Education_Level					
College Plus	15084 (65%)	2411 (100%)	12673 (100%)	0 (0%)	0 (0%)
No College	7951 (35%)	0 (0%)	0 (0%)	1710 (100%)	6241 (100%)
Part_and_Perf	0.05 (0.02, 0.24)	0.87 (0.60, 0.95)	0.04 (0.02, 0.09)	0.71 (0.52, 0.94)	0.04 (0.02, 0.11)
Percent_Grade	0.01 (0.00, 0.13)	0.87 (0.74, 0.93)	0.00 (0.00, 0.03)	0.84 (0.64, 0.92)	0.00 (0.00, 0.04)
event_count.x_logged	6.34 (5.34, 7.56)	8.32 (7.96, 8.71)	5.94 (5.14, 6.78)	8.39 (8.05, 8.73)	6.12 (5.19, 6.99)
Event_Count_total	564 (209, 1910)	4122 (2864, 6084)	381 (171, 881)	4384 (3133, 6191)	455 (179, 1084)
Relative_Grade_to_Engagement_Ratio	0.17 (0.00, 0.65)	0.98 (0.89, 1.13)	0.00 (0.00, 0.35)	0.99 (0.86, 1.61)	0.00 (0.00, 0.38)

[1] Statistics presented: n (%); median (IQR)

Note: College plus all-rounders (10.4%), college plus disengagers (55%), no college all-rounders (7.4%), no college disengagers (27.1%). Total N = 23,305, excluding all unknown education level data.

When unknown is included, six clusters are derived, two for each education background level, College Plus, No College, and Unknown. These conclusions are represented in Figure 5 and Table 6 from the original analysis.

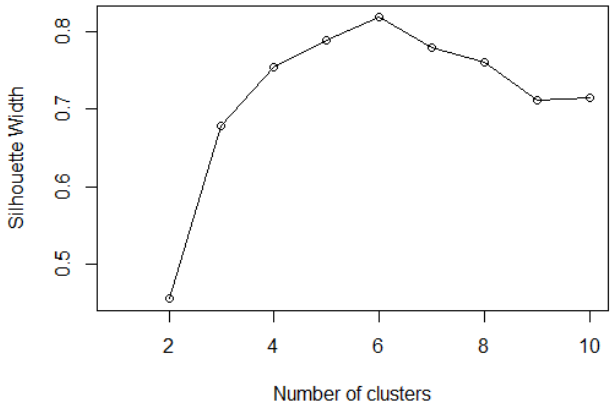

Figure 5: Silhouette plot of the PAM clustering algorithm for k=2:10. At six clusters, the average silhouette width for the data objects is above 0.8, indicating sound clusters. N = 29,083, including all unknown education level data

Table 6: Descriptive Statistics of the Six Clusters: PAM Algorithm

Characteristic	Overall, N = 29083	College Plus, All-rounders, N = 2411[1]	College Plus, Disengagers, N = 12673[1]	No College, All-rounders, N = 1710[1]	No College, Disengagers, N = 6241[1]	Unknown, All-rounders, N = 942[1]	Unknown, Disengagers, N = 5106[1]
Education_Level							
College Plus	15084 (52%)	2411 (100%)	12673 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
No College	7951 (27%)	0 (0%)	0 (0%)	1710 (100%)	6241 (100%)	0 (0%)	0 (0%)
Unknown	6048 (21%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	942 (100%)	5106 (100%)
Part_and_Perf	0.05 (0.02, 0.23)	0.87 (0.60, 0.95)	0.04 (0.02, 0.09)	0.71 (0.52, 0.94)	0.04 (0.02, 0.11)	0.78 (0.54, 0.92)	0.03 (0.02, 0.08)
Percent_Grade	0.01 (0.00, 0.11)	0.87 (0.74, 0.93)	0.00 (0.00, 0.03)	0.84 (0.64, 0.92)	0.00 (0.00, 0.04)	0.85 (0.66, 0.92)	0.00 (0.00, 0.01)
Event_Count_total	547 (202, 1850)	4122 (2864, 6084)	381 (171, 881)	4384 (3133, 6191)	455 (179, 1084)	4444 (3162, 6224)	366 (141, 849)
Relative_Grade_to_Engagement_Ratio	0.12 (0.00, 0.61)	0.98 (0.89, 1.13)	0.00 (0.00, 0.35)	0.99 (0.86, 1.61)	0.00 (0.00, 0.38)	0.99 (0.86, 1.48)	0.00 (0.00, 0.23)

[1] Statistics presented: n (%); median (IQR)

Note: College plus all-rounders (8.3%), college plus disengagers (44%), no college all-rounders (5.9%), no college disengagers (21%), unknown all-rounders (3.2%), and unknown disengagers (18%). Total N = 29,083.

The second important consideration stems from the inclusion of SES data. Specifically, only a small sample of SES data from the USA is available for analysis. This data is represented alongside the entire "committed learner" data set, including data from outside the USA. It could be potentially problematic to cluster on entire world data and then represent SES data from only the USA, especially if the clustering results would be different between entire world data and USA-only data. When limiting the Gower distance–based cluster analysis to USA data only, six clusters are found, extremely similar to the six clusters found across the full sample of "committed learners"; thus there was no need to differentiate the clustering results further.

Limiting the Gower distance–based cluster analysis data to USA data only, six clusters are found to be appropriate according to silhouette width analysis, represented in Figure A.2, yielding extremely similar clusters, represented in Table A.2.

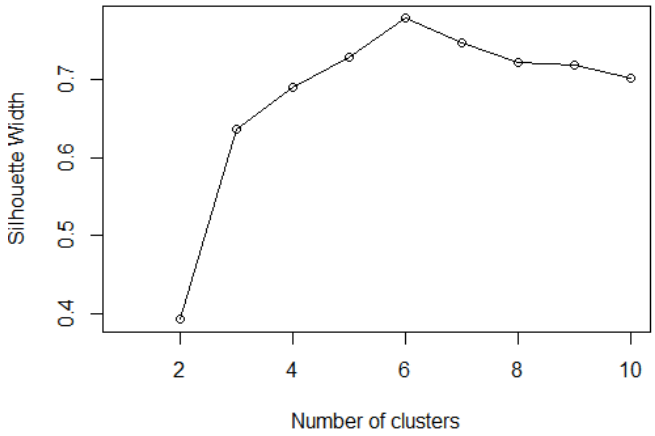

Figure A.2: Silhouette plot of the PAM clustering algorithm for k=2:10, USA-only data. At six clusters, the average silhouette width for the data objects is well above 0.7 indicating sound clusters. Total N = 9,708, including only USA data

Table A.2: Descriptive Statistics of the Six Clusters: PAM Algorithm

Characteristic	**Overall**, N = 9708	**College Plus, All-rounders**, N = 654[1]	**College Plus, Disengagers**, N = 2761[1]	**No College, All-rounders**, N = 1097[1]	**No College, Disengagers**, N = 2681[1]	**Unknown, All-rounders**, N = 541[1]	**Unknown, Disengagers**, N = 1974[1]
Education_Level							
College Plus	3415 (35%)	654 (100%)	2761 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
No College	3778 (39%)	0 (0%)	0 (0%)	1097 (100%)	2681 (100%)	0 (0%)	0 (0%)
Unknown	2515 (26%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	541 (100%)	1974 (100%)
Part_and_Perf	0.08 (0.02, 0.37)	0.83 (0.56, 0.95)	0.04 (0.02, 0.11)	0.59 (0.48, 0.90)	0.05 (0.02, 0.14)	0.76 (0.54, 0.92)	0.03 (0.02, 0.11)
Percent_Grade	0.02 (0.00, 0.30)	0.87 (0.72, 0.94)	0.00 (0.00, 0.04)	0.83 (0.63, 0.91)	0.01 (0.00, 0.05)	0.85 (0.69, 0.91)	0.00 (0.00, 0.03)
event_count.x_logged	6.66 (5.54, 7.90)	8.36 (7.97, 8.77)	6.03 (5.19, 6.94)	8.38 (8.04, 8.73)	6.34 (5.38, 7.17)	8.43 (8.11, 8.76)	6.07 (5.14, 6.91)
Event_Count_total	784 (254, 2704)	4264 (2890, 6468)	417 (179, 1033)	4350 (3115, 6202)	564 (216, 1300)	4561 (3328, 6371)	433 (170, 1001)
Relative_Grade_to_Engagement_Ratio	0.25 (0.00, 0.82)	0.99 (0.89, 1.31)	0.00 (0.00, 0.41)	1.03 (0.87, 1.69)	0.14 (0.00, 0.42)	1.00 (0.86, 1.58)	0.00 (0.00, 0.25)

[1] Statistics presented: n (%); median (IQR)

Note: College plus all-rounders (6.7%), college plus disengagers (28.4%), no college all-rounders (11.3%), no college disengagers (28%), unknown all-rounders (5.5%), and unknown disengagers (20.3%). Total N = 9,708, including only USA data.

When the entire data set is analyzed, six extremely similar clusters are derived, two for each education background level, College Plus, No College, and Unknown. These conclusions are represented in Figure 5 and Table 6 from the original analysis.

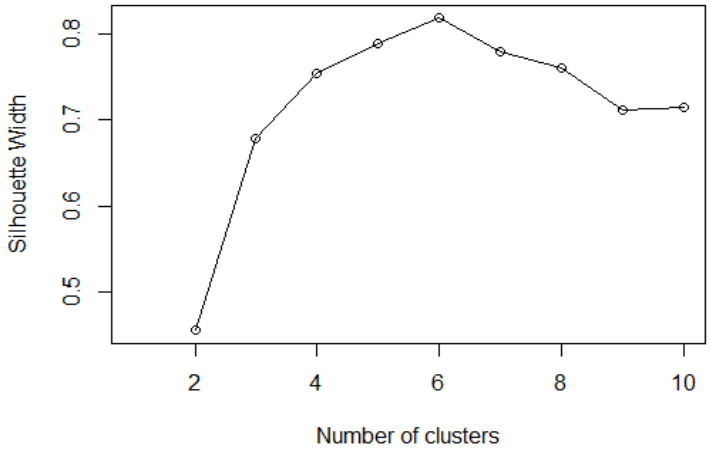

Figure 5: Silhouette plot of the PAM clustering algorithm for k=2:10. At six clusters, the average silhouette width for the data objects is above 0.8, indicating sound clusters. N = 29,083, including all unknown education level data.

Table 6: Descriptive Statistics of the Six Clusters: PAM Algorithm

Characteristic	**Overall**, N = 29083	**College Plus, All-rounders**, N = 2411[1]	**College Plus, Disengagers**, N = 12673[1]	**No College, All-rounders**, N = 1710[1]	**No College, Disengagers**, N = 6241[1]	**Unknown, All-rounders**, N = 942[1]	**Unknown, Disengagers**, N = 5106[1]
Education_Level							
College Plus	15084 (52%)	2411 (100%)	12673 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
No College	7951 (27%)	0 (0%)	0 (0%)	1710 (100%)	6241 (100%)	0 (0%)	0 (0%)
Unknown	6048 (21%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	942 (100%)	5106 (100%)
Part_and_Perf	0.05 (0.02, 0.23)	0.87 (0.60, 0.95)	0.04 (0.02, 0.09)	0.71 (0.52, 0.94)	0.04 (0.02, 0.11)	0.78 (0.54, 0.92)	0.03 (0.02, 0.08)
Percent_Grade	0.01 (0.00, 0.11)	0.87 (0.74, 0.93)	0.00 (0.00, 0.03)	0.84 (0.64, 0.92)	0.00 (0.00, 0.04)	0.85 (0.66, 0.92)	0.00 (0.00, 0.01)
Event_Count_total	547 (202, 1850)	4122 (2864, 6084)	381 (171, 881)	4384 (3133, 6191)	455 (179, 1084)	4444 (3162, 6224)	366 (141, 849)
Relative_Grade_to_Engagement_Ratio	0.12 (0.00, 0.61)	0.98 (0.89, 1.13)	0.00 (0.00, 0.35)	0.99 (0.86, 1.61)	0.00 (0.00, 0.38)	0.99 (0.86, 1.48)	0.00 (0.00, 0.23)

[1] Statistics presented: n (%); median (IQR)

Note: College plus all-rounders (8.3%), college plus disengagers (44%), no college all-rounders (5.9%), no college disengagers (21%), unknown all-rounders (3.2%), and unknown disengagers (18%). Total N = 29,083.

Finally, the distribution of learners with available SES background is similar across the USA-only clusters and the entire-world clusters. Table A.3 represents the multinomial logistic regression output of USA-only data where SES is the explanatory

variable, and cluster is the outcome variable. Table 8 is from the original analysis, indicating similar results; notably, that users from lower SES backgrounds are just as likely to be in any of the clusters compared to their higher SES peers.

Table A.3: Relative Risk Ratios: SES and Cluster

	Dependent variable:				
	College Plus, Disengagers	No College, All-rounders	No College, Disengagers	Unknown, All-rounders	Unknown, Disengagers
	(1)	(2)	(3)	(4)	(5)
bi_SES_status_finalUnknown SES	1.167	0.490***	0.818	0.262***	0.654***
	(0.124)	(0.131)	(0.122)	(0.144)	(0.124)
bi_SES_status_finalLow SES	0.934	1.213	1.255	1.470	1.447
	(0.285)	(0.286)	(0.272)	(0.293)	(0.273)
Constant	3.718***	2.864***	4.781***	2.041***	4.145***
	(0.115)	(0.119)	(0.112)	(0.125)	(0.114)
Akaike Inf. Crit.	31,279.030	31,279.030	31,279.030	31,279.030	31,279.030
Note:					*p<0.1; **p<0.05; ***p<0.01

Note: The relative risk ratios, shown as the exponentiated value of the logit coefficients from the multinomial logistic regression, where SES is the explanatory variable, and cluster is the outcome variable. College plus, all-rounders from mid-high SES backgrounds are the reference group. Total N = 9,708, including only USA data.

Table 8: Relative Risk Ratios: SES and Cluster

	Dependent variable:				
	College Plus, Disengagers	No College, All-rounders	No College, Disengagers	Unknown, All-rounders	Unknown, Disengagers
	(1)	(2)	(3)	(4)	(5)
bi_SES_status_finalUnknown SES	1.376***	0.229***	0.544***	0.160***	0.518***
	(0.111)	(0.119)	(0.110)	(0.128)	(0.112)
bi_SES_status_finalLow SES	0.909	1.257	1.259	1.471	1.428
	(0.270)	(0.274)	(0.260)	(0.282)	(0.261)
Constant	3.878***	2.613***	4.538***	1.877***	3.868***
	(0.109)	(0.114)	(0.107)	(0.120)	(0.109)
Akaike Inf. Crit.	85,200.670	85,200.670	85,200.670	85,200.670	85,200.670
Note:					*p<0.1; **p<0.05; ***p<0.01

Note: The relative risk ratios, shown as the exponentiated value of the logit coefficients from the multinomial logistic regression, where SES is the explanatory variable, and cluster is the outcome variable. College plus, all-rounders from mid-high SES backgrounds are the reference group. N = 29083.