Data-Driven Modeling of Learners' Individual Differences for Predicting Engagement and Success in Online Learning

Kamil Akhuseyinoglu
University of Pittsburgh
Pittsburgh, PA, USA
kaa108@pitt.edu

Peter Brusilovsky
University of Pittsburgh
Pittsburgh, PA, USA
peterb@pitt.edu

ABSTRACT

Individual differences have been recognized as an important factor in the learning process. However, there are few successes in using known dimensions of individual differences in solving an important problem of predicting student performance and engagement in online learning. At the same time, learning analytics research has demonstrated that the large volume of learning data collected by modern e-learning systems could be used to recognize student behavior patterns and could be used to connect these patterns with measures of student performance. Our paper attempts to bridge these two research directions. By applying a sequence mining approach to a large volume of learner data collected by an online learning system, we build models of student learning behavior. However, instead of following modern work on behavior mining (i.e., using this behavior directly for performance prediction tasks), we attempt to follow traditional work on modeling individual differences in quantifying this behavior on a latent data-driven personality scale. Our research shows that this data-driven model of individual differences performs significantly better than several traditional models of individual differences in predicting important parameters of the learning process, such as success and engagement.

CCS CONCEPTS

• **Social and professional topics** → **Computer science education**; • **Applied computing** → **Interactive learning environments**; • **Information systems** → **Structured Query Language**.

KEYWORDS

individual differences, learner modeling, sequential pattern mining, learning technology, online practice, SQL

ACM Reference Format:
Kamil Akhuseyinoglu and Peter Brusilovsky. 2021. Data-Driven Modeling of Learners' Individual Differences for Predicting Engagement and Success in Online Learning. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '21), June 21–25, 2021, Utrecht, Netherlands.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3450613.3456834

UMAP '21, June 21–25, 2021, Utrecht, Netherlands

ACM ISBN 978-1-4503-8366-0/21/06…$15.00
https://doi.org/10.1145/3450613.3456834

1 INTRODUCTION

Individual differences have been recognized as an important factor in the learning process. A wide range of cognitive, personal, motivational, and other dimensions of individual differences was introduced by researchers in the areas of cognitive science and educational psychology [36]. However, traditional dimensions of individual differences haven't yet proven their value in addressing the needs of modern e-learning. In particular, there are few successes in using these differences in solving the important problem of predicting student performance and engagement [1, 16, 56]. At the same time, e-learning research has demonstrated that the large volume of learning data collected by modern e-learning systems could be used to recognize student behavior patterns and connect these patterns with measures of student performance [8, 26, 27, 31, 41, 47, 52]. Our paper attempts to bridge these research directions. We use logs of student practice in an online practice system to identify patterns of student behavior and to reveal latent groups that exhibit considerably different practice behavior. We use these groups to model latent individual differences as a continuous behavior scale, similar to traditional models of individual differences, such as the achievement goal orientation framework and the self-esteem scale [23, 50]. In our study, this data-driven model of individual differences performed significantly better than traditional models of individual differences in predicting learner success and engagement.

2 RELATED WORK

2.1 Individual Differences and Academic Achievement

Individual differences have been the focus of research on educational psychology and learning technology [36]. Numerous works have attempted to discover and examine various *dimensions* of individual differences, find their connections to academic achievement, and address these differences in order to better support teaching and learning. A learner's position within a specific dimension of individual differences is usually determined by processing carefully calibrated questionnaires and placing the learner on a linear scale, frequently between two extreme ends. In this section, we briefly review several dimensions of individual differences that are frequently used in learning technology research.

Self-efficacy refers to one's evaluation of their ability to perform a future task [6] and is shown to be a good predictor of educational performance [12, 48]. Students with higher self-efficacy beliefs are more willing to put effort into learning tasks and persist more, as compared to students with lower self-efficacy. *Self-esteem* represents individuals' beliefs about their self-worth and competence [44]. Some studies have shown the positive effect of self-esteem

on academic achievement, while other studies have pointed out how academic achievement affects self-esteem [7, 20]. Researchers also stated the indirect effect of low self-esteem on achievement through distress and decreased motivation [40]. Learners can also differ by their *achievement goals*, which guide their learning behaviors and performance by defining the expectations used to evaluate success [38]. Studies have demonstrated the positive effects of achievement goals on performance [32, 39]. There are several known questionnaire-based instruments to capture achievement goals [23, 45].

Another important group of individual differences is related to metacognition, which plays an important role in academic performance [22]. In particular, students who successfully distinguish what they know and do not know can expand their knowledge instead of concentrating on already mastered concepts [55]. It has been shown that high-achieving students are more accurate in assessing their knowledge [21]. To measure some metacognitive differences, Tobias and Everson [54] proposed a *knowledge monitoring assessment* instrument to evaluate the discrepancy between the actual performance of students and their own estimates of their knowledge in a specific domain.

2.2 User Behavior Modeling and Performance Prediction

The rise of MOOCs has led to increased attention to learner data collected by MOOCs and similar online learning systems. The original motivation could be traced to the surprisingly high dropout rate of early MOOCs, which was hard to explain. Since MOOCs usually recorded full traces of learner behavior producing rich data for a large number of students, it was natural to use this data to predict dropouts [5] and performance [2, 15]. This appealing research direction quickly engaged researchers from the educational datamining community who were working on log mining and performance prediction in other educational contexts and led to a rapid expansion of research that connected learner behavior with learning outcomes in MOOCs and beyond.

While the first generation of this research focused on one-step MOOC performance prediction from learning data [2, 10, 11, 15], the second generation attempted to uncover the roots of performance differences to better understand the process and improve predictions. The core assumption of this stream of work was the presence of latent learner cohorts that exhibited similar behavior patterns and the connection of these patterns to performance and outcomes. While the idea of cohorts was pioneered by in the first generation research, the early work on cohorts attempted to define them using either learner demographic [30] or simple activity measures [2, 52]. In contrast, the second generation research attempted to automatically discover these cohorts from available data. Over just a few years, a range of approaches to discover behavior patterns and use them to cluster learners were explored. This included various combinations of clustering [9, 35], transition analysis [9, 27], Markov models [26, 31, 52], matrix factorization [41, 46, 47], tensor factorization [59], and sequence mining [8, 31, 35, 46, 58], which is reviewed in more detail in the next section.

2.3 Sequential Pattern Mining

In educational research, mining sequential patterns has become one of the common techniques to analyze and model students' activity sequences. This technique helped researchers to find student learning behaviors in different learning environments. Nesbit et al. [49] applied this technique to find self-regulated behaviors in a multimedia learning environment. In [42], authors identified the most frequent usage interactions to detect high/low performing students in collaborative learning activities. To find differences among predefined groups (e.g. high-performing/low-performing), Kinnebrew et al. [37] proposed a differential sequence mining procedure by analyzing the students' frequent patterns. Herold et al. [33] used sequential pattern mining to predict course performance, based on sequences of handwritten tasks. Guerra et al. [29] examined the students' problem solving patterns to detect stable and distinguishable student behaviors. In addition, Hosseini et al. [35] used a similar approach to [29] and detected different student coding behaviors on mandatory programming assignments, as well as their impact on student performance. Venant et al. [58] discovered frequent sequential patterns of students' learning actions in a laboratory environment and identified learning strategies that associated with learners' performance. Recently, Mirzaei et al. [46] explored specific patterns in learner behavior by applying both sequential pattern mining and matrix factorization approaches.

3 SYSTEM AND DATASET

We explored the prospects of data-driven modeling of individual differences by examining student behavior and learning outcomes in an online practice system for SQL programming. The system was available over several semesters to students taking a database class. The non-mandatory nature of the system allowed students to decide when and how much to practice and increased their chances to expose individual differences through their practice behavior. In addition to the logs of online practice, the dataset used for our study included data from pre-tests, post-tests, and several questionnaires that all focused on individual differences. This section explains in detail the nature of the practice system and components of the dataset. The next section (Section 4) explains how this original dataset was augmented with a new data-driven dimension of individual differences distilled from the log data.

3.1 The Course and the Online Practice System

In this study, we use data collected from four semesters of classroom studies in a graduate level Database Management course at a large North American university. Learning Structured Query Language (SQL) was one of the objectives of the course. The structure of the course remained the same for all four semesters, including the syllabus and the grading policy.

The SQL practice system [13] was offered to all classes as a non-mandatory tool for learning and self-assessment. The system provided access to two types of interactive learning content: SQL problems focused on SQL SELECT statements and annotated examples of SQL statements [14]. The content was grouped into topics, and each topic had multiple problems and examples. Students could choose the topic and the content to practice in any order. To encourage the students to explore the practice system, one percentage

point of extra credit was provided to the students who solved at least 10 SQL problems.

The problems were designed to help students practice their SQL code writing skills. Each problem was *parameterized*; i.e., generated from a set of pre-defined templates with randomly selected parameters. This design allowed students to practice the same problem multiple times. The correctness of their responses was tested against a fixed database schema and immediate correct/incorrect feedback was provided. The problem tool also had a unique feature, called *query execution mode*, which allows students to execute their SQL queries multiple times to see the actual query result tables while working with the problems. After checking their query results, students could submit their final query and get immediate feedback from the tool. As another content type, *annotated examples* offered worked examples of SQL code augmented with explanations, which students could examine interactively, line by line. The students practiced with 46 problem templates and 64 annotated examples with 268 distinct explanation lines.

3.2 The Dataset Collection

Knowledge Metrics: To measure overall knowledge improvement throughout the course, a *pre-test* and a *post-test* were administered, and *normalized learning gain* (NLG) was calculated as the ratio of the actual gain to the maximum possible gain:

$$NLG = (post - pre)/(max\_possible\_post - pre) \quad .$$

Each test had 10 questions that required writing SQL statements. Reported pre- and post-test scores ranged between 0 and 10.

Individual Differences: We collected gender data, and used several instruments to measure individual differences. To measure *global self-worth*, we used a 10-item Rosenberg Self-Esteem Inventory [50]. Responses ($\alpha = .82$) were converted to a continuous scale where higher scores indicate higher *self-esteem* (SE). To perform a *Tobias-Everson knowledge monitoring assessment* (KMA) [54], we asked students' estimates about their answers to pre-test problems. Then, based on correct and incorrect answers and estimates, a score was computed that ranges from -1 to 1, where a score of 1 indicates that the student knows perfectly what they know or do not know.

Activity Logs: We collected students' timed interaction logs with the online practice system for each of the four consecutive semesters. The logs offer a detailed view of student interaction, including each attempt to solve a problem and access to each example line. The data collected from the first three semesters were used for our data-driven behavioral modeling, and we refer to these three semesters as the *modeling dataset*. Interaction logs collected during semester 4 were used as a test dataset.

Throughout the paper, we used the term *incoming differences* to refer to the gender, pre-test scores, SE, and KMA scores collected at the start of the course. We consider post-test scores and NLG to be the performance measures. Table 1 presents the summary about the participation, incoming differences, performance, and practice system usage for each semester. In our dataset, no student took the course in multiple semesters. Also, note that some students didn't respond to instruments and pre/post tests. In this table and the remaining analysis, we only used the data collected from students who gave their consent and who tried the practice system by attempting at least one SQL problem and viewing at least one example. Detailed filtering process shared along with the reported analyses. For practice system usage, we reported the average number of attempted distinct problems, viewed distinct examples, and explanation lines, as well as the average number of query execution requests.

4 MODELING LATENT INDIVIDUAL DIFFERENCES FROM PRACTICE BEHAVIOR

The goal of the work presented in this paper is to build a data-driven model of individual differences by processing and understanding learners' behavior within the practice system. In essence, we wanted to augment various dimensions of incoming individual differences already present in our dataset (i.e., gender, pre-test scores, SE and KMA) with another dimension distilled from data, and to compare the value of these dimensions in predicting performance and engagement. Given the non-mandatory nature of the practice system, students accessed practice problems and examples without predefined order or deadlines. We expected that this freedom of access increased the chance for latent learning-related dimensions of individual differences to be exposed through practice behavior and captured by behavior modeling. Past success of behavior mining approaches based on sequence mining encouraged us to apply sequence mining to discover behavior patterns and to use it for modeling latent individual differences. A distinctive feature of our approach among other sequence-mining approaches is representing the behavior of individual learners as a stable vector of behavior micro-patterns. The micro-patterns are used to discover latent groups of learners with similar behaviors. By following traditional questionnaire-based approaches to model individual differences, we considered these latent groups as opposite points on a latent behavioral scale and attempted to position every learner on this scale. This section presents the main steps of our approach in detail.

4.1 Learning Action Labeling

The first step in sequential pattern mining is to label students' practice actions and define the specific action sequences to be mined. We believed that the sequence of interactions with learning activities and transitions between the activities (i.e., examples and problems) were critical in modeling individual differences. To pursue this idea, we performed a labeling process that highlights these critical interactions. We started the labeling process by mapping each student action to a unique label. Table 2 lists key learning actions and the corresponding labels used in the labeling process. As described earlier, practice activities were grouped into several SQL topics. To access a list of activities for a topic, a student opens a topic. Once the topic is opened, learners can work with activities of the topic in any order. With this design, student work with a topic becomes a unit of practice. To reflect this, we formed behavior sequences corresponding to learners' work with individual topics: all learning actions between two topic openings are considered to be one sequence, and each sequence starts with the topic opening label *topic-o*. We also introduced labels for opening and working with each type of content (i.e., *ex-o, ex-line*). If a student performed a content action after opening a content item (attempting a problem

Table 1: Summary statistics of the collected dataset. Mean and (SD) are reported.

		Incoming Differences				Performance		Practice System Usage			
Semester	N	Female	SE	KMA	Pre-test	Post-test	NLG	Problems	Examples	Lines	Query Executions
Semester 1	44	54%	20.7(4.2)	.5(.4)	1.1(1.1)	5.1(1.5)	.46(.13)	31.2(17.7)	50.4(19.6)	123.6(69.4)	55.6(60.7)
Semester 2	22	42%	21.2(4.7)	.7(.3)	2.2(2.1)	4.8(2.2)	.33(.21)	31.2(16.1)	51.7(14.3)	108.8(72.3)	53.4(66.5)
Semester 3	22	55%	22.6(3.3)	.4(.6)	.9(1.0)	4.6(1.7)	.41(.17)	39.1(12.6)	59.6(10.1)	134.3(70.9)	52.0(51.3)
Semester 4	36	NA	NA	NA	1.9(1.9)	5.2(2.2)	.40(.24)	33.0(17.4)	51.2(19.0)	132.6(65.5)	57.8(75.6)

Table 2: List of labels and the corresponding learning actions that were used in the labeling process.

Pattern Label	Learning Action
topic-o	Opening a topic.
prob-s	Successful problem solving attempt.
ex-o	Opening an example activity.
prob-f	Failed attempt for a problem.
ex-line	Viewing an explanation line.
query-o	Opening query execution mode.
prob-o	Opening a problem.
query-e	Checking query results in query execution mode.

or viewing an explanation line), we collapsed labels for content opening and kept the labels for the actual learning actions. For example, a sequence *{prob-o, prob-o}* means that a student opened two problems consecutively without trying to solve any of them. In addition, we distinguished a failed and a successful problem solving attempt from one another to differentiate learning actions that occurred after either a failed or a successful attempt.

One of the challenges of sequence analysis of learning data is the presence of repetitive learning actions, such as a row of failed problem solving attempts, or a row of multiple line explanation views where exact number or repetition is not essential, but the relative scale of repetition is. To address it, we collapsed these sequences so that we can capture what actually happened after these repetitive actions. In this process, we first generated all sequences with repetitive labels. Then, we calculated the median length of repeated labels for all students. Then, we went over the original action sequences and replaced each repetitive label with a single uppercase version of that label if the length of that repetition was greater than the median length, or with a single lowercase label otherwise. At the end of this process, each label could represent one or more consecutive repeated actions, depending on the median length. Only *ex-line* and *query-e* had a median length of two, while others had a median length of one. As the result of the labeling process, 3432 sequences were generated from interaction logs of 88 students in the modeling dataset.

4.2 Sequential Pattern Mining

To discover the frequent patterns in student action sequences, we used the SPAM sequence mining algorithm [4, 24]. The sequences generated after the activity labeling process were used for mining frequent patterns. To reveal sequences that could highlight individual differences, we set the minimum support for the SPAM algorithm at 0.5% . Due to our labeling process with repetition reduction, the sequences used in the mining process were already dense in information. Even if some sequences were not frequently followed (not having high levels of support), they could be important in revealing discriminative practice behaviors. The SPAM algorithm discovered 169 frequent patterns that appeared at least in 0.5% of sequences (18 sequences). All discovered patterns consist of two or three consecutive learning actions, as we did not include any gap constraint to the SPAM algorithm. Table 3 shows the top 5 most frequent patterns.

Table 3: Discovered top 5 frequent patterns with sequence explanations and frequency of occurrence. Lowercase actions mean that the repetition of that action is less than or equal to the median repetition length, while uppercase actions mean the opposite.

Pattern	Freq.	Explanation
{topic-o, EX-LINE}	4.8%	Opening a topic followed by viewing a *long* sequence of line explanations.
{topic-o, ex-line}	2.7%	Opening a topic followed by viewing a *short* sequence of line explanations.
{topic-o, prob-o}	2.5%	Opening a topic followed by problem openings without any attempt.
{prob-f, query-e}	2.3%	Failed attempt followed by query executions.
{topic-o, EX-O}	2.2%	Opening a topic followed by a long sequence of example openings without line viewing.

Out of 88 students, 82 students had at least one frequent pattern after the mining process. We further filtered out students with less than 25 frequent patterns (Q1: 45.75, Med: 97.00, M: 103.20) to have a fair amount of representation of practice behavior by the discovered frequent patterns. After the filtering process, the number of students with frequent patterns dropped to 75.

4.3 Modeling Individual Practice Behavior with Frequent Patterns

We built an individual behavior profile for each student as a frequency vector using the discovered 169 frequent patterns. Each position in this vector represents how many times the corresponding frequent pattern appears in the practice work of the modeled student. To eliminate any possible impact of the amount of practice, we normalized the frequency vectors per student and now the resulting vectors represent the probability of the occurrence of

each frequent pattern. This approach was first introduced in [29] and successfully used to model learner behavior in [35]. Following these works, we called the individual behavior profile the *practice genome*.

To make sure that the constructed probability vectors represented a sufficiently stable profile of individual practice and could reliably distinguish students from each other, we checked the *stability* of the *practice genome*, as suggested in [29]. Following the suggested procedure, we split students' sequences into two halves using two approaches: (1) by random split, and (2) by temporal split. In the random-split approach, we shuffled students' topic-level sequences randomly and divided them into halves. In the temporal-split approach, we first ordered the sequences based on time and divided the sequences into early and late halves. For either split approach, we built separate practice "half-profiles" from each of the halves and calculated the pairwise distances for the whole set of 'half-profiles" using the Jenson-Shannon (JS) divergence (as we are calculating the distance between two probability distributions). To assert profile stability, the distance between the two "half-profile" vectors of the same student (self-distance) should be smaller than the distance to half-vectors of other students (others-distance). If this expectation holds, then we would find strong empirical evidence that the practice profiles (genomes) do represent some stable individual behavior that distinguishes a particular student from others.

To evaluate this expectation, we conducted a paired t-test to compare the calculated *self-distances* to *others-distances* for both random-split and temporal-split approaches. The random-split self-distances ($M = 0.35, SD = 0.11$) were significantly smaller than the random-split other-distances ($M = 0.46, SD = 0.05$); $t(79) = -7.531, p < .001$. Similarly, the temporal-split self-distances ($M = 0.42, SD = 0.11$) were significantly smaller than the temporal-split other-distances ($M = 0.48, SD = 0.05$); $t(76) = -5.034, p < .001$. These findings showed that the *practice genomes* constructed with the frequent patterns were stable in representing students' practice behavior and were successful in distinguishing students from each other. This property opens a way to use genomes for modeling individual differences.

4.4 Discovering Latent Groups of Learners Based on Practice Behavior

Given the stability of individual genome profiles, our next step was to discover behavioral clusters that group students with similar behavior profiles. The clustering was performed in two steps. First, we mapped the higher-dimensional practice genomes (i.e., 169 dimensions of the probability vectors) into a two-dimensional space by using a dimensionality-reduction technique. Next, we clustered students using the lower-dimensional representation of the practice profiles. The main rationale behind following a two-step clustering approach was that the low-dimensional representation enabled us to convert categorical cluster representation into a continuous behavioral scale, as explained in Section 4.7. In our approach, we fixed the number of clusters to two ($k = 2$) by analyzing the higher-dimensional practice genomes using silhouette method [51] and gap statistics [53].

During the first step of the clustering process, we used t-Stochastic Neighbor Embedding (t-SNE) [57], a non-linear dimensionality-reduction algorithm that is mainly used to explore high-dimensional data, to project practice genomes to 2-D points. t-SNE minimizes the objective function using a randomly-initiated gradient descent optimization. Thus, each run of t-SNE generates a different projection. For the results presented in this paper, we first applied a grid-search technique to tune hyper-parameters (e.g., exaggeration factor, perplexity, theta) and selected the projection that leads to the most distinct cluster separation (in Step 2), based on the frequent patterns. Thus, for the grid-search and the projection selection, we executed the first and the second step of the clustering process together for each run.

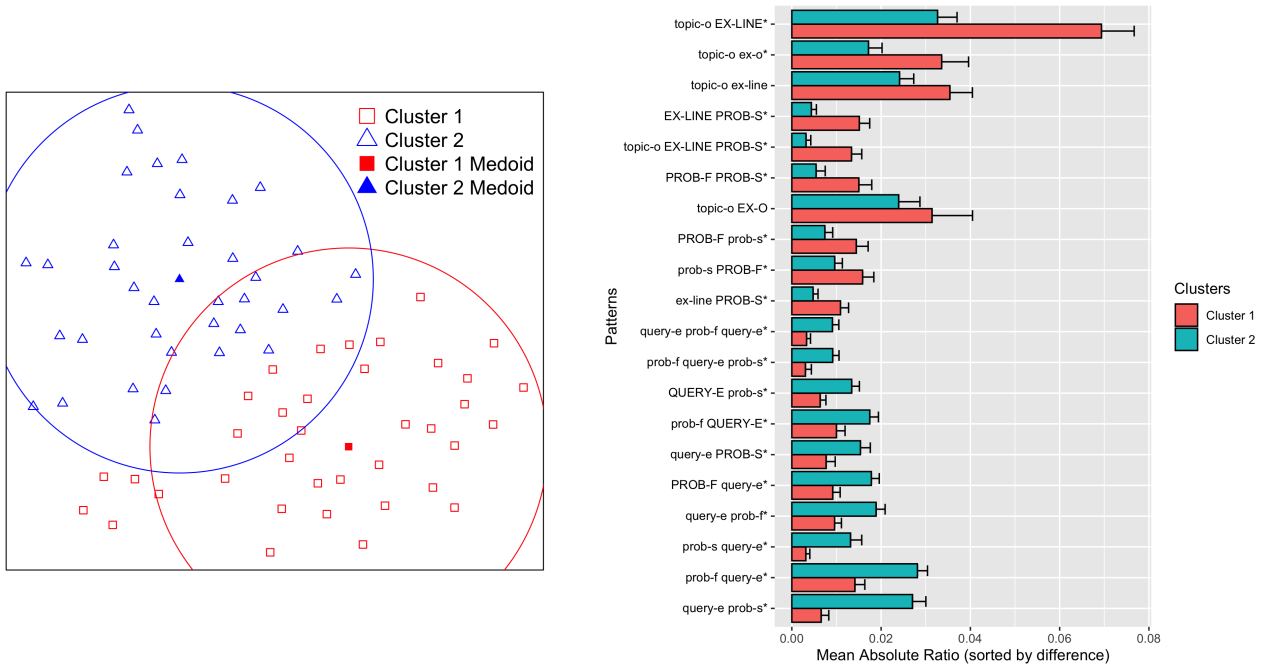
During the second step of the clustering process, we applied partition around medoids (PAM) clustering to the 2-D results of t-SNE projections. To judge the cluster separation for the grid-search and projection selection, we performed a differential sequence mining approach [37] to compare the mean probability (ratio) of each frequent pattern between the discovered clusters ($k = 2$) using multiple t-tests at $\alpha = 0.05$ and counted the number of significantly different patterns between each cluster. Based on this approach, we selected the 2-D t-SNE projection. The selected t-SNE projection of the practice behaviors and the PAM clustering results are presented in Figure 1a. After clustering, there were 38 and 37 students in clusters 1 and 2, respectively.

4.5 Practice Behaviors Discovered by Clusters

In this section, we review behavioral differences between the clusters by comparing frequent patterns exhibited by students in each cluster. To achieve this, we calculated the average ratio (probability) of frequent patterns in both clusters. In Figure 1b, we plotted the average ratio of 20 patterns that had the highest absolute ratio difference between two clusters and sorted them by the difference of the absolute ratio. In the figure, there are 10 patterns that more frequently occurred in Cluster 1 (top half of the the graph) and 10 patterns that more frequently occurred in Cluster 2 (bottom half of the graph). The significantly different patterns are labeled with a star.

As shown in the figure, students in two clusters exhibited considerably different practice behavior. Students in Cluster 1 significantly more frequently opened and explored examples right after they began to work with a topic (e.g.,{topic-o, EX-LINE}, {topic-o, ex-o}). Moreover, they switched more frequently from viewing explanations to successful problem solving, suggesting that they valued examples as a preparation tool for problem solving (e.g., {EX-LINE, PROB-S}, {topic-o, EX-LINE, PROB-S}). The students in this cluster were also engaged significantly more frequently in a sequence of uninterrupted problem solving attempts, in which a sequence of failed attempts was followed by a sequence of successful attempts (e.g., {PROB-F, PROB-S}, {PROB-F, prob-s}).

In contrast, students in Cluster 2 interleaved attempts to solve problems by using the query execution mode. As seen in the figure, all 10 "distinguishing" patterns that involved the query execution mode (e.g., query-e) were significantly more frequent in Cluster 2. For example, when Cluster 2 students failed on a problem, they checked their query results in the query execution mode to get more



(a) (b)

Figure 1: (a) Student practice behavior representation on 2-D t-SNE projection with PAM clustering results (k=2). The red circle represents the mean Euclidean distance of all students to the Cluster 1 medoid (the filled red square), and the blue circle represents the mean Euclidean distance of all students to the Cluster 2 medoid (the filled blue triangle). (b) Frequent pattern comparison between clusters, sorted by mean ratio (probability) difference. Stars denote significantly different patterns between two clusters, based on the results of the t-test.

detailed feedback (e.g., {prob-f, query-e}, {PROB-F,query-e}). Further, they typically managed to solve problems after using the query execution mode (e.g., {query-e, prob-s}, {query-e, PROB-S}, {prob-f, query-e, prob-s}. In some cases, they used the query execution mode even after successfully solving a problem (e.g., {prob-s, query-e}), suggesting that at particular cases they wanted to verify their correct queries by checking the actual query result.

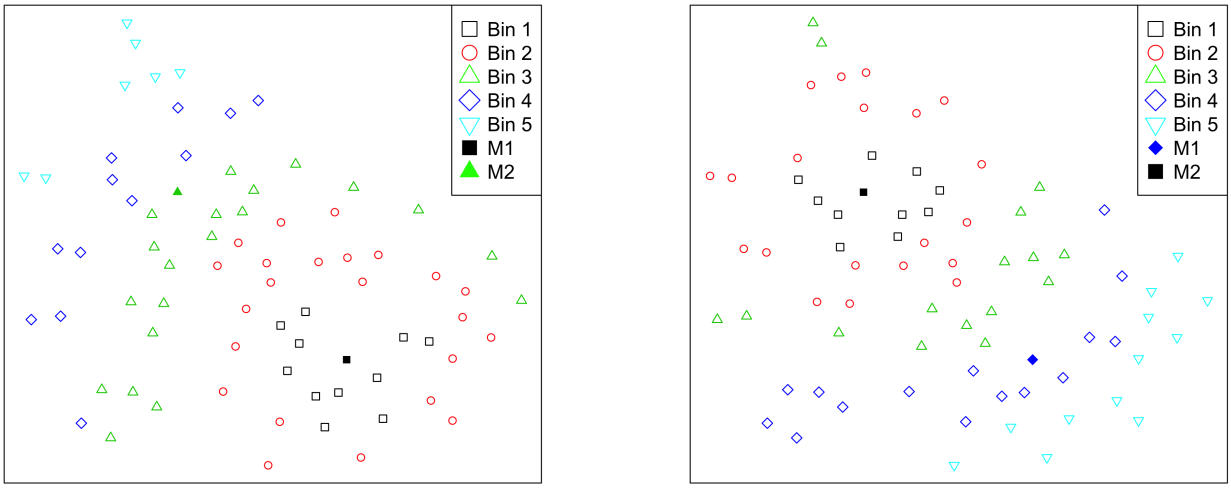
Table 4: Summary of the incoming differences and performance for the clusters. Mean and (SD) are reported.

		Incoming Differences				Performance	
Clusters	N	Female	SE	KMA	Pre-test	Post-test	NLG
Cluster 1	38	56%	21.2(4.2)	.53(.44)	1.2(1.5)	5.2 (1.6)	.45(.14)
Cluster 2	37	44%	21.7(4.4)	.60(.45)	1.4(1.6)	4.8 (1.9)	.39(.20)

In summary, by clustering practice genomes, we discovered two divergent practice behaviors. To simplify the difference, Cluster 1 students tended to learn by consuming SQL knowledge encapsulated in examples and then applying it to practice problems. Cluster 2 students preferred to "generate" SQL knowledge through their own experience obtained by experimenting with various SQL queries, which they used as exploration, debugging, and verification tools.

4.6 Connecting Clusters to Incoming Differences

While the discovered clusters revealed some divergent learning behavior, it was important to check whether the observed differences could be explained through already collected incoming individual differences (i.e., gender, pre-test scores, SE, and KMA). If this connection could not be established, we could hypothesize that the observed differences reflect some latent dimension of individual differences that could be used to construct the new scale. To connect the obtained clusters to incoming differences, we checked whether there were any other noticeable differences between clusters in terms of individual differences or prior knowledge. We summarized these comparisons in Table 4. For each measure, we only considered students with available data. As seen in the table, the clusters were balanced according to the incoming differences. The percentage of female students were 56% and 44%, respectively. The mean scores of SE, KMA, and pre-test were also very similar, and we did not find any statistically significant differences. We also did not find any significant differences between clusters based on post-test scores and NLG. Further, using pre-test, KMA, SE scores, and gender, we fitted a binomial generalized linear model to predict categorical cluster assignments. Compared to an intercept-only model that used the likelihood ratio test, incoming differences did not improve the overall fit of the model ($\chi^2(3) = 1.383, p = .710$) and achieved a



(a) (b)

Figure 2: Students are divided into five bins based on their Euclidean distances to cluster medoids. Bins are numbered from 1 to 5 in increasing average distance for (a) Cluster 1 medoid (M1) and (b) Cluster 2 medoid (M2).

very low area under the ROC curve (AUC) of 0.587, which suggests that cluster assignments cannot be fully explained by the incoming differences.

4.7 Developing a Data-Driven Behavioral Scale

In the last step of the behavior modeling process, we attempted to refine the categorical cluster assignments into a continuous behavioral scale (metric) that can model individual differences reflected in the practice behavior, similar to traditional scales of individual differences. For example, the SE scale is not a categorical scale (e.g., high/low SE), but rather a continuous representation of global self-worth. We believed that a binary categorization simplifies the observed variability in practice behavior.

To follow existing "bi-polar" scales of individual differences, we attempted to quantify the position of a student with respect to each main practice behavior (depicted by the clusters) as the Euclidean distance from the student's 2-D point to the cluster medoids found by the PAM clustering algorithm. Thus, we modelled the practice behavior of a student using two numerical values: (1) distance to the first cluster medoid (M1), and (2) distance to the second cluster medoid (M2).

To investigate how distances to cluster medoids captured differences among students (i.e., incoming differences, engagement, and performance), we divided students into five *bins* using distances. The bins are numbered from 1 to 5 in increasing average distance for M1 and M2, where bin 1 is the closest group to the medoids, as illustrated in Figure 2. As Table 5 shows, grouping based on distance to M1, the average number of distinct explanation lines viewed drops considerably as the distance increases, and we found a significant negative correlation ($r = -.48, p < .001$). We also found a weak positive correlation with the average number of query executions ($r = 0.23, p = .04$), but there is no constant decrease or increase regarding this usage metric. For other usage metrics, incoming differences, and performance measures, we did not find any significant correlation with the distance to M1. Thus, with the increase of the distance to M1, the number of distinct line views decreases and the number of query executions increases. For grouping based on the distance to M2, we found a significant negative correlation between distance and the number of distinct problems attempted ($r = -.30, p = .009$), and between distance and the number of query executions ($r = -68, p < .001$). We also found a significant positive correlation with the NLG ($r = 0.24, p = .038$). Thus, we can summarize that when students move away from M2, the NLG increases while they attempted fewer distinct problems and performed fewer number of query executions.

The correlations summarized in this section overlaps with the main behavioral patterns described in Section 4.5, where students in Cluster 1 were more concentrated on examples and students in Cluster 2 were performing more query executions. It is important to highlight that we did not find any significant correlation between distance and the incoming differences. Using pre-test scores, KMA, SE scores and gender, we predicted distance values by fitting a linear regression model. However, we did not find any significant predictive model. This means that we cannot explain distances to cluster medoids and the attributed practice behavior by the incoming differences.

Table 5: Summary of grouping based on distances to the cluster medoids. Mean values are reported.

		Incoming Differences				Performance		Practice System Usage			
Dist. to M1	N	Female	SE	KMA	Pre-test	Post-test	NLG	Problems	Examples	Lines	Query Exec.
Bin 1	12	40%	21.80	.46	1.17	5.40	.49	41.50	60.50	170.00	51.25
Bin 2	23	55%	21.65	.57	1.37	4.96	.41	37.00	56.13	152.39	35.30
Bin 3	21	56%	21.68	.65	1.24	4.78	.41	35.76	56.47	141.86	90.14
Bin 4	12	42%	20.17	.50	1.17	4.83	.41	41.08	56.00	93.75	81.92
Bin 5	7	50%	21.67	.59	1.71	5.17	.39	35.28	52.43	61.57	61.57
Dist. to M2	N	Female%	SE	KMA	Pre-test	Post-test	NLG	Problems	Examples	Lines	Query Exec.
Bin 1	11	44%	22.27	.59	.45	4.27	.40	42.18	58.73	156.72	153.36
Bin 2	20	53%	21.89	.59	1.82	4.73	.34	42.35	57.55	124.80	86.65
Bin 3	16	43%	19.93	.51	.81	5.00	.46	36.81	55.19	123.81	38.31
Bin 4	16	43%	21.00	.63	2.06	5.80	.48	34.63	56.31	125.13	31.50
Bin 5	12	70%	21.17	.49	.83	4.96	.45	32.17	55.08	156.42	16.42

5 PREDICTING ENGAGEMENT AND PERFORMANCE

In this section, we evaluate the predictive power of the continuous behavioral metric on various engagement and performance measures. Performance represents the outcomes of the learning process and is the most typical measure of the learning process. Engagement is a less traditional group of metrics, yet it remains essential in the context of non-mandatory learner-driven practice. While practicing with interactive learning content is usually beneficial for the growth of learner knowledge, many students tend to ignore the opportunity to practice or practice very little. In this context, the engagement with practice becomes a critical parameter of the learning process. The ability to connect individual differences with engagement is important to predict the outcomes of the learning process and to plan interventions.

We compared the relative predictive power of the behavioral metric (i.e., distance to cluster medoids) against the incoming individual differences (i.e., gender, pre-test scores, KMA, and SE scores) and the categorical behavioral cluster representations. Further, we checked the transferability of the constructed behavioral metric using the test dataset.

To perform these comparisons, we fitted multiple regression models to separately predict each outcome measure and compared the overall fit of models using likelihood ratio tests. Moreover, we compared the relative importance of features based on the regression estimates. We used negative binomial generalized linear models to predict count outcome variables due to over-dispersion. For other measures, we fitted simple linear regressions. We considered adding a random effect to the regression models to account for the variability in semesters, but given the very low estimated variance of the random effect, we continued with only the fixed effects models. In addition, we checked regression assumptions, including the multicollinearity, by calculating the *variance inflation factors (VIF)* and made sure that none of the features had $\sqrt{VIF} > 2$.

5.1 Performance and Engagement Measures

Learner engagement, an important factor in the learning process, has been extensively discussed in the research literature [3, 28]. In modern e-learning, engagement is frequently approximated by the amount of student voluntary work (i.e., work not directly required and graded). For example, in MOOCs, engagement is frequently assessed by the fraction of watched videos, the number of attempted quizzes, or the number of posts to a discussion forum [2, 17, 18]. Similarly, online practice systems generally measure learner engagement through the amount of voluntary practice with examples and problems [19, 34]. Following this practice, we approximated *engagement* as the amount of students' non-mandatory work with different learning activities available in the practice system. The measures that we used for engagement are: (1) the total number of learning actions performed calculated by summing up the number of problem solving attempts (regardless of correctness), the number of query execution attempts, the number of annotated examples viewed, and the number of line explanations viewed (referred as *total-actions*); and (2) the total number of distinct learning activities: problems, annotated examples, and line explanations (referred as *dist-content*). Note that the *total-actions* measure counts duplicate accesses to the same learning content, such as opening the same example or attempting to solve the same problem more than once. Thus, this metric reflects the overall levels of engagement with the practice system. On the other hand, *dist-content* incorporates uniqueness of the learning content and reflects overall content coverage by a student. To measure student *performance*, we used (3) post-test scores and (4) NLG as the objective performance metrics collected outside the practice system.

5.2 Predicting Engagement

For engagement prediction analyses, we used 70 students' data from the modeling dataset who filled out the survey, took the pre-test, attempted at least one problem, and viewed at least one example in the practice system. Table 6 presents the summary of the fitted models for engagement measures.

We started our analyses by predicting the total number of learning actions performed in the practice system (*total-actions*). Results indicated that the incoming individual differences did not improve the overall fit of the model when compared to an intercept-only model($\chi^2(4) = 5.230, p = .265$). However, the model with distance to medoids as features (distance-model) fitted data significantly better than both the intercept-only model ($\chi^2(2) = 14.744, p < .001$) and the model that used the incoming differences (incoming-model) as features ($\chi^2(2) = 9.514, p = .009$). Moreover, the distance-model also fitted data significantly better than the model that used categorical cluster labels (cluster-model) ($\chi^2(1) = 12.191, p = .001$). We further fitted a separate model with both distance measures and individual differences together as features (full-model) to compare relative regression coefficients. In this model, distance to M1 and distance to M2 were the only significant predictors of total-actions, where moving away from both medoids was associated with a fewer number of actions. We can conclude that the distance to M2 had a slightly higher negative effect on *total-actions*.

As a second task, we fitted models to predict the total number of distinctly accessed learning contents (*dist-content*). Similar to the results for *total-actions*, none of the student individual differences significantly predicted *dist-content* and did not improve the overall fit of the model, as compared to an intercept-only model ($\chi^2(4) = 4.146, p = .387$). On the other hand, we found out that the distance-model fitted the engagement data significantly better than the *intercept-only* model ($\chi^2(2) = 10.575, p = .005$), the *cluster-model* ($\chi^2(1) = 10.557, p = .001$), and the *incoming-model* ($\chi^2(2) = 6.428, p = .004$). Next, we fitted a model with both distance and incoming differences together (full-model) and found that only the distance measures were significant predictors. Here, we saw that the distance to M1 had a higher negative impact on *dist-content*, as compared to the distance to M2, which was based on regression coefficients.

In summary, students who were close to M2 performed more total actions, which can be explained by the overall practice behavior of Cluster 2: they failed more on problem attempts and used query execution more frequently, as compared to Cluster 1. Students who were close to M1 covered more unique content in total, such as viewing more explanation lines.

5.3 Predicting Performance

In this section, we advanced to learning outcome prediction by predicting both post-test scores and the NLG. For performance prediction analyses, we used the same set of students in the engagement prediction, but excluded four students who did not take the post-test and who had zero learning gain. Finally, we used 66 students' data for learning outcome prediction.

In learning outcome prediction, we used the same features that we explored in engagement prediction. Moreover, we wanted to control for the students' "amount of practice" by adding system usage metrics to our regression models as additional features. We considered the total number of distinct problems attempted, distinct examples viewed, and distinct explanation lines viewed as possible features. We performed a backward step-wise feature selection process and found out that the total number of *distinct problems attempted*(DPA) was the only feature that significantly predicts the post-test scores (after controlling by the pre-test scores) and NLG. Thus, we added this usage metric as a feature to all our regression models. The summary of the fitted models is presented in Table 6.

To predict post-test scores, in addition to the DPA feature, we added pre-test scores to control for the levels of prior knowledge. First, we fitted a regression model by adding remaining features for incoming differences; i.e., KMA and SE scores (incoming-model). Compared to a regression model with pre-test scores and DPA as features, the incoming-model did not fitted the data better ($\chi^2(3) = 0.326, p = .955$) and none of the features related to incoming differences were significant, except for the pre-test scores ($B = 0.620, t = 5.824, p < .001$) and DPA ($B = 0.216, t = 2.128, p = .037$). On the other hand, we replicated the similar analysis by fitting a model with distance to medoids as features (distance-model), and found out that the distance-model fitted the data significantly better than the incoming-model ($\chi^2(1) = 4.928, p = .026$), and better than the model that added binary cluster assignments as a feature to pre-test scores and DPA ($\chi^2(1) = 4.121, p = .042$). Next, we fitted a model with all features together (full-model), and these results indicated that after pre-test scores and DPA, the distance to M2 was a significant predictor, but that the distance to M1 was not. Thus, after controlling for the prior knowledge and the number of distinct problems attempted, the distance to the second cluster medoid significantly predicts post-test scores.

In predicting NLG, we did not include pre-test scores as a feature, since it was used to calculate NLG. Similar to post-test prediction results, in separate models, we discovered that none of the incoming differences and binary cluster labels were significant predictors. However, the model with distance features (distance-model) showed that distance to M2 significantly predicted NLG after controlling for the DPA. We further fitted a model with all features together (full-model), and again, only DPA and distance to M2 were significant predictors. Given the positive sign of the M2 regression coefficients in both post-test and NLG predictions, we concluded that the distance from the Cluster 2 medoid is associated with higher learning performance.

5.4 Transferability of Learning Outcome Prediction

In this section, we assess the transferability of our behavior modeling approach by predicting the learning outcomes of students in a test dataset that was not used to discover the latent groups and build the behavior scale. In addition, to assess whether the behavior modeling approach can be used for the practical needs of predicting student performance well before the course is finished, we only used students' action sequences from the first half of the course. There were 36 students who used the practice system (attempted at least one problem and viewed at least one example) in the test dataset. We filtered out students who did not take both pre- and post-tests and who did not have any frequent patterns that could be used to build the practice genome. After this filtering process, 27 students remained.

The main challenge in this process was projecting new students' practice genomes on an already constructed 2-D tSNE projection, as shown in Figure 1a, because the t-SNE algorithm learns a non-parametric mapping. To overcome this challenge, we trained a

Table 6: Summary of fitted regression models to predict engagement and performance measures in the modeling dataset.

	Dependent variable:							
	total-actions		dist-content		Post-test scores		NLG	
	distance-model	*full-model*	*distance-model*	*full-model*	*distance-model*	*full-model*	*distance-model*	*full-model*
Pre-test scores		−.002		.0001	.611***	.587***		
Gender (M)		.124		.050		.112		.258
SE score		.097		.071		−.031		−.075
KMA score		−.071		−.053		−.012		−.019
DPA					.293***	.291***	.354***	.353***
Dist. to M1	−.216***	−.207***	−.201***	−.196***	.124	.124	.194	.184
Dist. to M2	−.292***	−.289***	−.121**	−.121**	.264**	.266**	.365**	.361**
Adjusted R^2					.435	.409	.104	.083
Log Likelihood	-485.570	-483.257	-412.689	-410.900				
Akaike Inf. Crit.	977.141	980.513	831.378	835.801				
F Statistic					13.486***	7.434***	3.527**	1.984*

*p<0.1; **p<0.05; ***p<0.01

multivariate regression model to predict the location (x and y coordinates) of new practice genomes on a 2-D map using the practice genomes from the modeling dataset. This way, we can predict new students' locations and calculate their distances to the same cluster medoids.

Using the first-half sequences, we discovered 109 frequent patterns following the same approach presented in Section 4, where 102 of the patterns overlapped with the previously discovered patterns and 49 of them overlapped with the previously discovered top-50 frequent patterns in the modeling dataset. To build the model and not overfit the modeling dataset, we only used these overlapping frequent patterns as features and further reduced this set to 28 patterns by applying a multivariate backward step-wise feature selection procedure. The final trained model explained the variance in the coordinates reasonably well (x: $adj.R^2 = 0.89$, y: $adj.R^2 = 0.83$) and convinced us to proceed. Using this model, we predicted the locations of new students on the 2-D map and calculated Euclidean distances to both medoids.

Similar to the analyses in Section 5.3 to predict post-test scores, in addition to the DPA feature, we added pre-test scores to control for levels of prior knowledge. Since we did not have individual incoming difference measures for these students, we can only report prediction results of the distance measures on this new dataset. Our results indicated that the overall model was significant ($F(4, 22) = 5.365, p = .004, adj.R^2 = 0.40$). Compared to the same model fitted in the modeling dataset, we lost 3.4% in explained variance (based on adj. R^2, 0.435 in modeling dataset and 0.401 in test dataset), but this finding could simply be a result of using only half-sequences of students. Based on the regression results, similar to our previous findings, the distance to M2 was a significant predictor($B = 0.642, t = 2.588, p = .017$) but not the distance to M1 ($B = -0.357, t = -1.312, p = .203$). In NLG prediction, we found a similar trend where the distance to M2 was a significant feature ($B = 0.605, t = 2.072, p = .049$) but not the distance to M1 ($B = -0.525, t = -1.758, p = .092$). These results indicate that the model of practice behavior that was built by using the modeling dataset represents a reasonably stable dimension of individual differences that could be used in new datasets to predict learning outcomes.

6 DISCUSSIONS AND CONCLUSION

In this paper, we proposed a data-driven approach to reveal and model latent individual differences in online practice behavior. Using three semesters of log data from an online practice system, we revealed latent clusters of learners with different behavior and converted categorical cluster assignments into a continuous scale representing individual differences in practice behavior. We evaluated this scale against the original dataset and examined the transferability of our modeling approach against a new semester-long dataset. Our findings showed that the data-driven behavioral metric can predict both learners' engagement within the online practice system and their learning outcomes. In contrast, categorical cluster assignments were not equally successful in predicting the overall levels of performance and engagement. This data suggests that the discovered "practice behavior" should be modeled on a continuous scale.

Our results showed that "closeness" to one of the cluster medoids was associated with higher learning outcomes. However, we obtained this result after controlling for the practice "efforts". This finding indicates that learning outcomes are not defined solely by the sheer amount of practice efforts, but also by how a student practiced.

The reported results are interesting and important, but our study does have limitations. The first group of limitations is related to the measures applied. Based on multiple regression analysis, we showed that traditionally modelled individual differences, such as self-esteem (SE) and knowledge monitoring assessment (KMA), were not effective in both engagement and performance prediction. This finding is supporting the prior research showing that SE has no impact on performance [7]. However, the SE measure used in this study focuses on global self-worth. More specific *self-concept* constructs had a stronger relationship to traditional levels of academic achievement [43]. In addition, we administered KMA on SQL problems that required students to write short SQL statements without having any options to select. We believe that the nature of

such problems might reduce the predictive power of KMA. Other modified versions of similar measures could be used [25] or knowledge assessment could be monitored during usage of the practice system. In addition, the performance measures that we used in this paper were based on pre/post tests rather than on actual course grades, due to having no access to this information.

Since the practice system offered as a non-mandatory resource, our analyses are subject to self-selection bias, where we can only observe the practice behavior of students who decided to use the system. The design of the practice system adds another limitation to our findings, where students had freedom to choose topics and content on which to work freely. In addition, we collected the data from similar student cohorts attending the same graduate-level course at a large North American university. The results presented in this paper might not be transferable to other cohorts or cultures. Similar studies and analyses should be conducted in other courses and in other cultures in different settings to assess the generality of the study results. Finally, we only reported associations between behavior and learning outcomes, not claiming any causality. In our future work, we plan to apply this modeling approach to a newer version of the practice system that has more types of learning content, explore cultural differences, incorporate other self-reported measures, and check the differences among practice behaviors in the presence of different engagement manipulations.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 1740775.

REFERENCES

[1] Donggun An and Martha Carr. 2017. Learning styles theory fails to explain learning and achievement: Recommendations for alternative approaches. *Personality and Individual Differences* 116 (2017), 410 – 416.

[2] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2014. Engaging with Massive Online Courses. In *Proceedings of the 23rd International Conference on World Wide Web* (Seoul, Korea) *(WWW '14)*. Association for Computing Machinery, New York, NY, USA, 687–698.

[3] James J Appleton, Sandra L Christenson, Dongjin Kim, and Amy L Reschly. 2006. Measuring cognitive and psychological engagement: Validation of the Student Engagement Instrument. *Journal of School Psychology* 44, 5 (2006), 427–445.

[4] Jay Ayres, Jason Flannick, Johannes Gehrke, and Tomi Yiu. 2002. Sequential pattern mining using a bitmap representation. In *Proceedings of the Eight ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '02)*. ACM, 429–435.

[5] Girish Balakrishnan. 2013. *Predicting Student Retention in Massive Open Online Courses using Hidden Markov Models*. Master's thesis. EECS Department, University of California, Berkeley. http://www2.eecs.berkeley.edu/Pubs/TechRpts/2013/EECS-2013-109.html

[6] Albert Bandura. 1982. Self-efficacy mechanism in human agency. *American Psychologist* 37, 2 (1982), 122.

[7] Roy F. Baumeister, Jennifer D. Campbell, Joachim I. Krueger, and Kathleen D. Vohs. 2003. Does High Self-Esteem Cause Better Performance, Interpersonal Success, Happiness, or Healthier Lifestyles? *Psychological Science in the Public Interest* 4, 1 (2003), 1–44.

[8] Mina Shirvani Boroujeni and Pierre Dillenbourg. 2018. Discovery and temporal analysis of latent study patterns in MOOC interaction sequences. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge (LAK '18)*. ACM, 206–215.

[9] Ahcène Boubekki, Shailee Jain, and Ulf Brefeld. 2018. Mining User Trajectories in Electronic Text Books. In *Proceedings of the 11th International Conference on Educational Data Mining (EDM 2018)*. 147–156.

[10] Sebastien Boyer and Kalyan Veeramachaneni. 2015. Transfer Learning for Predictive Models in Massive Open Online Courses. In *17th International Conference on Artificial Intelligence in Education, AIED 2015*. Madrid, Spain, 54–63.

[11] Christopher G. Brinton and Mung Chiang. 2015. MOOC performance prediction via clickstream data and social learning networks. In *2015 IEEE Conference on Computer Communications (INFOCOM)*. 7218617, 2299–2307.

[12] Shari L. Britner and Frank Pajares. 2006. Sources of science self-efficacy beliefs of middle school students. *Journal of Research in Science Teaching* 43, 5 (2006), 485–499.

[13] Peter Brusilovsky, Sibel Somyürek, Julio Guerra, Roya Hosseini, Vladimir Zadorozhny, and Paula J Durlach. 2016. Open Social Student Modeling for Personalized Learning. *IEEE Transactions on Emerging Topics in Computing* 4, 3 (2016), 450–461.

[14] Peter Brusilovsky, Sergey Sosnovsky, Michael V Yudelson, Danielle H Lee, Vladimir Zadorozhny, and Xin Zhou. 2010. Learning SQL programming with interactive tools: from integration to personalization. *ACM Transactions on Computing Education (TOCE)* 9, 4 (2010), 19.

[15] John Champaign, Kimberly Colvin, Alwina Liu, Colin Fredericks, Daniel Seaton, and David Pritchard. 2014. Correlating Skill and Improvement in 2 MOOCs with a Student's Time on Tasks. In *Proceedings of the First ACM Conference on Learning @ Scale Conference*. ACM, 11–20.

[16] Guanliang Chen, Dan Davis, Claudia Hauff, and Geert-Jan Houben. 2016. On the Impact of Personality in Massive Open Online Learning. In *Proceedings of the 24th Conference on User Modeling, Adaptation and Personalization (UMAP 2016)*. Association for Computing Machinery, 121–130.

[17] R. Wes Crues, Nigel Bosch, Michelle Perry, Lawrence Angrave, Najmuddin Shaik, and Suma Bhat. 2018. Refocusing the Lens on Engagement in MOOCs. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale* (London, United Kingdom) *(L@S '18)*. Association for Computing Machinery, New York, NY, USA, Article 11, 10 pages.

[18] Dan Davis, Ioana Jivet, René F. Kizilcec, Guanliang Chen, Claudia Hauff, and Geert-Jan Houben. 2017. Follow the Successful Crowd: Raising MOOC Completion Rates through Social Comparison at Scale. In *Proceedings of the Seventh International Learning Analytics and Knowledge Conference* (Vancouver, British Columbia, Canada) *(LAK '17)*. Association for Computing Machinery, New York, NY, USA, 454–463.

[19] Paul Denny, Fiona McDonald, Ruth Empson, Philip Kelly, and Andrew Petersen. 2018. *Empirical Support for a Causal Relationship Between Gamification and Learning Outcomes*. Association for Computing Machinery, New York, NY, USA, 1–13.

[20] Laura Di Giunta, Guido Alessandri, Maria Gerbino, Paula Luengo Kanacri, Antonio Zuffiano, and Gian Vittorio Caprara. 2013. The determinants of scholastic achievement: The contribution of personality traits, self-esteem, and academic self-efficacy. *Learning and Individual Differences* 27 (2013), 102 – 108.

[21] Daniell DiFrancesca, John L Nietfeld, and Li Cao. 2016. A comparison of high and low achieving students on self-regulated learning variables. *Learning and Individual Differences* 45 (2016), 228–236.

[22] David Dunning, Kerri Johnson, Joyce Ehrlinger, and Justin Kruger. 2003. Why People Fail to Recognize Their Own Incompetence. *Current Directions in Psychological Science* 12, 3 (2003), 83–87.

[23] Andrew J Elliot and Holly A McGregor. 2001. A 2×2 achievement goal framework. *Journal of Personality and Social Psychology* 80, 3 (2001), 501–519.

[24] Philippe Fournier-Viger, Jerry Chun-Wei Lin, Antonio Gomariz, Ted Gueniche, Azadeh Soltani, Zhihong Deng, and Hoang Thanh Lam. 2016. The SPMF open-source data mining library version 2. In *Joint European conference on machine learning and knowledge discovery in databases*. 36–40.

[25] Claudia Gama. 2004. Metacognition in Interactive Learning Environments: The Reflection Assistant Model. In *Proceedings of International Conference on Intelligent Tutoring Systems*. Springer Berlin Heidelberg, Berlin, Heidelberg, 668–677.

[26] Chase Geigle and ChengXiang Zhai. 2017. Modeling Student Behavior With Two-Layer Hidden Markov Models. *Journal of Educational Data Mining* 9, 1 (2017), 1–24.

[27] Niki Gitinabard, Sarah Heckman, Tiffany Barnes, and Collin F. Lynch. 2019. What will you do next? A Sequence Analysis of the Student Transitions between Online Platforms. In *Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)*. 59–68.

[28] Tabitha L Grier-Reed, J. Appleton, M. Rodriguez, Zoila M. Ganuza, and Amy L. Reschly. 2012. Exploring the Student Engagement Instrument and Career Perceptions with College Students. *Journal of Educational and Developmental Psychology* 2 (2012), 85–96.

[29] Julio Guerra, Shaghayegh Sahebi, Yu-Ru Lin, and Peter Brusilovsky. 2014. The Problem Solving Genome: Analyzing Sequential Patterns of Student Work with Parameterized Exercises. In *Proceedings of the 7th International Conference on Educational Data Mining (EDM 2014)*. 153–160.

[30] Philip Guo and Katharina Reinecke. 2014. Demographic Differences in How Students Navigate Through MOOCs. In *Proceedings of the First ACM Conference on Learning @ Scale Conference*. ACM, 21–30.

[31] Christian Hansen, Casper Hansen, Niklas Hjuler, Stephen Alstrup, and Christina Lioma. 2017. Sequence Modelling For Analysing Student Interaction with Educational Systems. In *Proceedings of the 10th International Conference on Educational Data Mining (EDM 2017)*. 232–237.

[32] Judith M Harackiewicz, Kenneth E Barron, Paul R Pintrich, Andrew J Elliot, and Todd M Thrash. 2002. Revision of achievement goal theory: Necessary and

illuminating. *Journal of Educational Psychology* 94, 3 (2002), 638–645.
[33] James Herold, Alex Zundel, and Thomas F Stahovich. 2013. Mining Meaningful Patterns from Students' Handwritten Coursework. In *Proceedings of the 6th International Conference on Educational Data Mining (EDM 2013)*. 67–73.
[34] Roya Hosseini, Kamil Akhuseyinoglu, Peter Brusilovsky, Lauri Malmi, Kerttu Pollari-Malmi, Christian Schunn, and Teemu Sirkiä. 2020. Improving Engagement in Program Construction Examples for Learning Python Programming. *International Journal of Artificial Intelligence in Education* 30, 2 (2020), 299–336.
[35] Roya Hosseini, Peter Brusilovsky, Michael Yudelson, and Arto Hellas. 2017. Stereotype Modeling for Problem-Solving Performance Predictions in MOOCs and Traditional Courses. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization (UMAP 2017)*. ACM, 76–84.
[36] David H. Jonassen and Barbara L. Grabowski. 1993. *Handbook of Individual Differences, Learning, and Instruction*. Routledge, New York, NY.
[37] John S Kinnebrew and Gautam Biswas. 2012. Identifying Learning Behaviors by Contextualizing Differential Sequence Mining with Action Features and Performance Evolution. In *Proceedings of the 5th International Conference on Educational Data Mining (EDM 2012)*. 57–64.
[38] Elizabeth A Linnenbrink and Paul R Pintrich. 2001. Multiple goals, multiple contexts: The dynamic interplay between personal goals and contextual goal stresses. In *Motivation in learning contexts: Theoretical advances and methodological implications*. Pergamon Press, Elmsford, NY, US, 251–269.
[39] Elizabeth A. Linnenbrink and Paul R. Pintrich. 2002. Achievement Goal Theory and Affect: An Asymmetrical Bidirectional Model. *Educational Psychologist* 37, 2 (2002), 69–78.
[40] Xiaoru Liu, Howard B Kaplan, and Will Risser. 1992. Decomposing the Reciprocal Relationships between Academic Achievement and General Self-Esteem. *Youth & Society* 24, 2 (1992), 123–148.
[41] Stephan Lorenzen, Niklas Hjuler, and Stephen Alstrup. 2018. Tracking Behavioral Patterns among Students in an Online Educational System. In *Proceedings of the 11th International Conference on Educational Data Mining (EDM 2018)*. 280–285.
[42] Roberto Martinez Maldonado, Kalina Yacef, Judy Kay, Ahmed Kharrufa, and Ammar Al-Qaraghuli. 2011. Analysing frequent sequential patterns of collaborative learning activity around an interactive tabletop. In *Proceedings of the 4th International Conference on Educational Data Mining (EDM 2011)*. 211–216.
[43] Herbert W. Marsh and Rhonda G. Craven. 2006. Reciprocal Effects of Self-Concept and Performance From a Multidimensional Perspective: Beyond Seductive Pleasure and Unidimensional Perspectives. *Perspectives on Psychological Science* 1, 2 (2006), 133–163.
[44] Gerald Matthews, Ian J Deary, and Martha C Whiteman. 2003. *Personality traits*. Cambridge University Press.
[45] Carol Midgley, Avi Kaplan, Michael Middleton, Martin L. Maehr, Tim Urdan, Lynley Hicks Anderman, Eric Anderman, and Robert Roeser. 1998. The Development and Validation of Scales Assessing Students' Achievement Goal Orientations. *Contemporary Educational Psychology* 23, 2 (1998), 113 – 131.
[46] Mehrdad Mirzaei, Shaghayegh Sahebi, and Peter Brusilovsky. 2020. Detecting Trait versus Performance Student Behavioral Patterns Using Discriminative Non-Negative Matrix Factorization. In *Proceedings of the 33rd International Florida Artificial Intelligence Research Society Conference*. 439–444.
[47] Kousuke Mouri, Atsushi Shimada, Chengjiu Yin, and Keiichi Kaneko. 2018. Discovering Hidden Browsing Patterns Using Non-Negative Matrix Factorization. In *Proceedings of the 11th International Conference on Educational Data Mining (EDM 2018)*. 568–571.
[48] Karen D Multon, Steven D Brown, and Robert W Lent. 1991. Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation. *Journal of Counseling Psychology* 38, 1 (1991), 30.
[49] John C Nesbit, Mingming Zhou, Yabo Xu, and P Winne. 2007. Advancing log analysis of student interactions with cognitive tools. In *12th biennial conference of the european association for research on learning and insruction (EARLI)*. 2–20.
[50] Morris Rosenberg. 1965. *Society and the adolescent self-image*. Princeton university press.
[51] Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* 20 (1987), 53–65.
[52] Kshitij Sharma, Patrick Jermann, and Pierre Dillenbourg. 2015. Identifying Styles and Paths toward Success in MOOCs. In *Proceedings of the 8th International Conference on Educational Data Mining (EDM 2015)*. 408–411.
[53] Robert Tibshirani, Guenther Walther, and Trevor Hastie. 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 (2001), 411–423.
[54] Sigmund Tobias and Howard Everson. 1996. Assessing Metacognitive Knowledge Monitoring. Report No. 96-01. *College Entrance Examination Board* (1996).
[55] Sigmund Tobias and Howard T Everson. 2009. The importance of knowing what you know: A knowledge monitoring framework for studying metacognition in education. In *Handbook of metacognition in education*. Routledge/Taylor & Francis Group, New York, NY, US, 107–127.
[56] Dominic Upton and Sally Adams. 2006. Individual Differences in Online Learning. *Psychology Learning & Teaching* 5, 2 (Sept. 2006), 141–145.
[57] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9, 86 (2008), 2579–2605.
[58] Rémi Venant, Kshitij Sharma, Philippe Vidal, Pierre Dillenbourg, and Julien Broisin. 2017. Using Sequential Pattern Mining to Explore Learners' Behaviors and Evaluate Their Correlation with Performance in Inquiry-Based Learning. In *Proceedings of the 12th European Conference on Technology Enhanced Learning (EC-TEL 2017)*. Springer, Cham, 286–299.
[59] Xidao Wen, Yu-Ru Lin, Xi Liu, Peter Brusilovsky, and Jordan Barria Pineda. 2019. Iterative Discriminant Tensor Factorization for Behavior Comparison in Massive Open Online Courses. In *The World Wide Web Conference (WWW '19)*. Association for Computing Machinery, 2068–2079.