From Detail to Context: Modeling Distributed Practice Intensity and Timing by Multiresolution Signal Analysis

Cheng-Yu Chung
Computer Science
Arizona State University
Cheng.Yu.Chung@asu.edu

I-Han Hsiao
Computer Science
Arizona State University
Sharon.Hsiao@asu.edu

ABSTRACT

The distributed practice effect suggests that students retain learning content better when they pace their practice over time. The key factors are practice dosage (intensity) and timing (when to practice and how in between). Inspired by the thriving development of image recognition, this study adopts one of the successful techniques, multiresolution analysis (MRA), to model distributed and spaced practice (SP). We consider a sequence of practice sessions as a signal of the student's learning strategy. Then, we apply the stationary wavelet transform (SWT) to extract practice patterns spaced by three periods: small, medium, large. The result reveals a positive correlation between the small-spaced practice and the exam grade. The benchmark against baseline feature models shows that the SP patterns significantly improve the goodness-of-fit and complements the baseline models. This work successfully demonstrates 1) the use of MRA in modeling sequential patterns by event intensity and event timing; 2) the MRA approach can be used as an alternative method to improve existing student models of practice effort.

Keywords

distributed practice effect, testing effect, stationary wavelet transforms, signal multiresolution analysis, feature extraction

1. INTRODUCTION

In the midst of blended and distance learning environments, it is increasingly important for students to manage their time efficiently. Numerous researchers have proposed and developed various student models to capture how students utilize their time during the learning process. The results have shown that distributed practice is a simple but effective time-management strategy for learning [5]. Essentially, distributed practice comprises the testing and spacing effects, which suggest that the retention of information increases when the learner practices retrieving it in multiple spaced-out practice sessions [3].

Optimal distributed practice requires a combination of both the *intensity* and the *timing* of the practice events. In other words, an expressive student model must capture the intensity of practice sessions spaced by different periods. Although the two features appear to be straightforward, it is not easy to incorporate them in a sequential behavior model. For example, typical sequence analysis or sequential pattern mining would expect discrete input data and extract common patterns in the data according to the sequence support (the number of occurrences). Finding a meaningful and interpretable threshold is usually an ad-hoc process and particularly challenging [4]. A great threshold value may increase the chance of losing detail, and a small value may introduce more noises and miss the context. In the case of distributed practice, when the practice sessions are far apart, such a frequency-based approach will require more data to ensure sufficient within- and between-sequences support for a pattern of interest. To address this modeling challenge, we are motivated to explore an alternative computational method to capture the detail as well as the context, which can capture both the intensity and the timing of events at the same time.

We rationalize that a student's practice sessions distributed over a timeline resemble a signal to her/his learning process where the strength of learning is quantified as the increasing or decreasing values about the occurrences of the underlying events. With this definition, we can utilize a signal processing tool to extract the structural variation which approximates distributed practice patterns. In this work, we adopt the stationary wavelet transform (SWT) algorithm for this purpose. SWT is a widely-used signal processing tool in an application such as image pattern recognition. The algorithm decomposes an input signal into multiple components and represents the original signal by information at different resolutions. With the emphasis on the structure, we believe that SWT will allow us to overcome the challenge where the amount of sequential data may not be big enough to maintain the sequence support. Additionally, applying SWT as a feature extraction method also allows us to examine structural nuances in behavior sequences.

2. RELATED WORK

2.1 Sequence Analysis in Educational Data Mining

A *behavior sequence* is a chronicle of an activity. It describes a collection of events, and the order of them is meaningful. We can choose different features to characterize such a sequence, e.g., types of events, arrangements of events, time gaps between events. The features directly affect what we can find out from the analysis. Sequence analysis, in general, can refer to any data model that involves a kind of behavior sequences its characterization. Extensive research in EDM has been using behavior sequence analysis to model students' development of knowledge or skills.

The most intuitive approach is sequential pattern mining, which aims to discover repeated string patterns, alignments, or the very next possible items [11]. For example, Gitinabard et al. characterize behavior sequences by students' interactions with online tools [7]. They map the interaction sequences to study habits and use sequence patterns to differentiate the high-performing and low-performing students. Dermy and Brun argue that the time interval is the key to model students' activities [4]. They characterize behavior sequences by time intervals between events and formalize the temporal information in sequential pattern mining. Their experiment suggests a strong correlation between the students' activities and the time information.

One research gap we notice is that most of the reviewed works focus on the behavior sequences at a single time scale. For example, for a given behavior sequence $e_1, e_2, ..., e_t$ where e_i is an event that occurs at time i. A typical sequence analysis focuses on the relationship of adjacent events e_{j-1}, e_j, e_{j+1} where $j \in 1, ..., t$. Since the step size is 1, sometimes such a sequence is called 1-sequence. Following this setting, a pattern must be a consecutive 1-sequences that meets predefined criteria, e.g., the support. One limitation of 1-sequences is that they cannot capture an inconsecutive event. Such an inconsecutive event can provide a coarser view of the behavior sequence, therefore the context. Indeed, we can try to increase the step size to have 2-sequences, 3-sequences, or k-sequences where $k \in \mathbb{Z}$. Nonetheless, the increment of step size inevitably reduces the number of k-sequences we can find in a dataset. This situation may exclude potential sequences of interest due to the threshold of the support or the shortage of data. To tackle this challenge, we investigate an alternative model that focuses on the structural information of behavior sequences.

3. MULTIRESOLUTION SIGNAL ANALYSIS

In pattern recognition, the information of a given object usually is determined by the variations of signal intensity. For example, we can recognize a building as a building in an image because the distinct contours and shapes are formulated by their unique signal value sequences and different from the other objects. Such signal *features* are essentially sequences of values (sets of numbers) where a variation of intensity could suggest a potential event of interest, e.g., a change of shapes or colors. However, because the objects to analyze may have different shapes and sizes, the feature extraction must consider "how far away" an event is from its neighborhood to recognize the objects' structures at multiple resolutions. The field of computer vision and signal processing have developed various methods to address this challenge. One of which is the multiresolution analysis and wavelet transforms, which fit in the scope of this research.

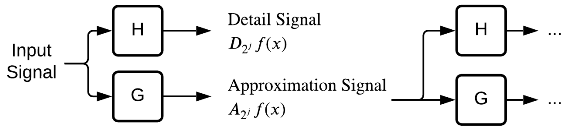

Figure 1: The Decomposition of Multiresolution Analysis. The process consists of two filters: the high-pass filter H and the low-pass filter G. They iteratively extract the detail signal and the approximation signal at the resolution 2^j from the input signal $f(x)$ until a maximum level L. We can associate the interpretation of the detail signal to the underlying time scale. For example, say the sampling rate of the input signal is 1. The detail signal at level 1 (a coarser level) denotes the information from the frequency band $[1/2, 1/4]$.

The multiresolution analysis (MRA) is a hierarchical framework that describes how to decompose a signal from fine to coarse levels [12]. The decomposition consists of a high-pass filter (H) and a low-pass filter (G). They are a pair of quadrature mirror filters and have the following relationship: $g(n) = (-1)^{1-n}h(1-n)$ [12]. The high-pass filter extracts impulses, and meanwhile, the low-pass one retains the other information. This process is also known as Discrete Wavelet Transform (DWT). By convolution ($*$), the filtering process iteratively produces series of detail signals ($D_{2^j} f(x)$) and approximation signals ($A_{2^j} f(x)$) for the input signal $f(x)$:

$$D_{2^j} f = (f(u) * \psi_{2^j}(-u))(2^{-j}n) \tag{1}$$

$$A_{2^j} f = (f(u) * \phi_{2^j}(-u))(2^{-j}n) \tag{2}$$

where $n \in \mathbb{Z}$. The high-pass and low-pass filters rely on a wavelet function (ψ) and a scaling function (ϕ) that translate and scale the input signal at different resolutions, respectively. We illustrate the whole filtering process in Figure 1 for reference. See [2] and [12] for more details about the math properties of the wavelet function and the scaling function.

3.1 Analyzing Distributed Practice via Signals

In this study, we focus on the practical implication of MRA and illustrate how it can help identify students' distributed practice patterns. Students, especially those in an online learning or a blended learning environment, usually have greater flexibility in self-pacing their studies. In other words, they can watch the lecture videos and practice quiz questions anytime at their convenience. This nature makes it challenging to analyze their behaviors on the timeline.

For example, in scenario A, when a semester is two to three months long, we may find out that the students' practice sessions are sparse and do not follow one unified schedule. This makes the time of sessions less discriminating in finding common behavioral patterns. Thus, the researcher may choose to ignore the time feature. An alternative approach (scenario B) is aggregating the practice sessions by a priori assumption (e.g., students always study on a week-by-week basis or right before a deadline). However, all the above approaches may inevitably lose some detail about how exactly the students utilize their schedules, either missing discriminating patterns over time (scenario A) or limited to those strictly abiding by the class-paced schedule (scenario B).

To model students' distributed practice behavior over time, his/her behavior can be denoted as a sequence of the events, f with T discrete time steps: $f = \{e_1, e_2, ..., e_T\}$ where e_i for $1 \leq i \leq T$ can be any activity event of our interest. A student distributes his/her practice sessions at different rates or frequencies according to his/her preferences, path, or pace. This representation is like a signal and enables the feasibility to apply a signal processing algorithm.

Similar to the pattern recognition in computer vision, a student's practice sessions are like the shapes and colors that may evolve according to the sequences of signals. The sessions may have different sizes, i.e., time gaps between any two sessions. In other words, we aim to extract *distributed spaced practice* (SP) that are subsets of the input behavior sequence: $SP_k \subseteq f$ where $k \in \mathbb{N}$ and any two consecutive event items $\{e_i, e_j\} \subset SP_k$ are spaced by k time steps. Following this idea, MRA is used to extract such a "feature" from sequences of practice sessions, and thereby interpret the output as distributed practice patterns. For a practice signal at sampling rate = 1/day, the output signals can represent the information at coarser rates, e.g., 1 per 4 days and 1 per 8 days.

3.2 Stationary Wavelet Transform

The output of DWT are signals that represent information at different resolutions (or frequencies). The typical implementation of DWT keeps downsampling the input signal to obtain the detail signal and the approximation signal at each resolution [12]. Therefore, the transform is time-variant. The detail signal at one level is a half shorter than the one at the previous level. This property may cause a misalignment in time/frequency, which will make the decomposition generate fewer feature values for analysis. In this study, we follow an alternative implementation of MRA, Stationary Wavelet Transform (SWT), which is time-invariant. SWT replaces the downsampling by upsampling at each step [6]. Research has shown that SWT can improve the approximation and a preferred approach for applications like breakdown point detection and denoising [1].

4. DATASETS

To evaluate the method, we use two semesters' datasets from the same undergraduate class offered in a four-year university in the United States: Spring 2018 (SP18) and Fall 2018 (FA18). Both sessions lasted about 3-month. The class was a typical lecture-style in-person class with weekly assignments and monthly exams. The two sessions were practically identical, having exactly the same syllabus, same instructor, same teaching assistants, except for minor adjustments to the exam questions. Note most students in FA18 shared a similar background in engineering because the class was a required class for first-year engineering students. In SP18, there were more students from non-engineering schools, which resulted in much more diverse student background.

An online practice platform was introduced to the students at the beginning and available throughout the semester. On the platform, students could take multiple-choice questions to practice and review the class content. For any given practice question, the students had unlimited chances to retry; for any attempt, the corrective feedback (correct answer) would be provided upon submission. The questions served like so-called "tasks" in the context of tutoring systems [16]. Each of the tasks aims to help the student master some knowledge (or embedded knowledge components). However, the practice activity is different from working with assignments: there is no "hard deadline" by which the students must complete the practice questions. The students can practice on the platform as a kind of self-assessment [13]. In other words, the activity is "self-paced" [18] and aligned with the actions of reviewing slides, taking quizzes, or other practices that students can do for their benefit whenever they want.

The students' practice activities were logged as transactions of events, including the timestamps, the questions, and the correctness of the attempts. We processed and transformed the data into sequences of daily practice intensity. Here, the term "intensity" refers to the number of unique questions solved by a student. Each day is assumed to be a complete practice session. The sequence of daily intensity thereby resembles a discrete-time signal sampled at a constant rate equal to 1 sample per day. We excluded some students' data from the analysis due to low usage (those who only had only one practice session throughout the semester). An overview of the datasets is described in Table 1.

In this study, the exam letter grade is used as the students' learning performance index. The exam letter grade ranges from A ($M \geq 90$), B ($80 \leq M < 90$), to C/D/F ($M < 80$) where M is the raw average of three exam scores.

5. REPRESENTING DISTRIBUTED PRACTICE BY SWT SIGNALS

There are several parameters required for our model pipeline: the wavelet for SWT, the padding scheme, the maximum decomposition level, and the penalty of change point detection. The Haar wavelet is adopted in the SWT algorithm implementation, due to the simplest form of wavelet [14]. It creates a shape like a step function that produces 1, 0, and -1, following the formula

$$\psi(x) = \begin{cases} 1 & \text{if } 0 \leq x < \frac{1}{2} \\ -1 & \text{if } \frac{1}{2} \leq x < 1 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

This property makes it a good option for detecting edges (e.g., sudden signal transitions or changes) [17] in discrete signals like the datasets in this study. The implementation of SWT used in this study requires the length of input to be a multiple of 2^L where L is the maximum number of levels to decompose [9]. To meet this requirement, we preprocessed all input sequences by adding a prefix of zeros. In our experiment, we found that the SWT signals at L > 3 did not work. It was likely due to short input sequences. Therefore,

Dataset	# of Students	# of Included	Max Length of Sequence (Days)	# of Questions	M (SD) of Intensity
SP18	121	76 (63%)	93	96	0.26 (0.36)
FA18	200	67 (34%)	82	95	0.32 (0.49)

Table 1: Statistics of the Two Datasets

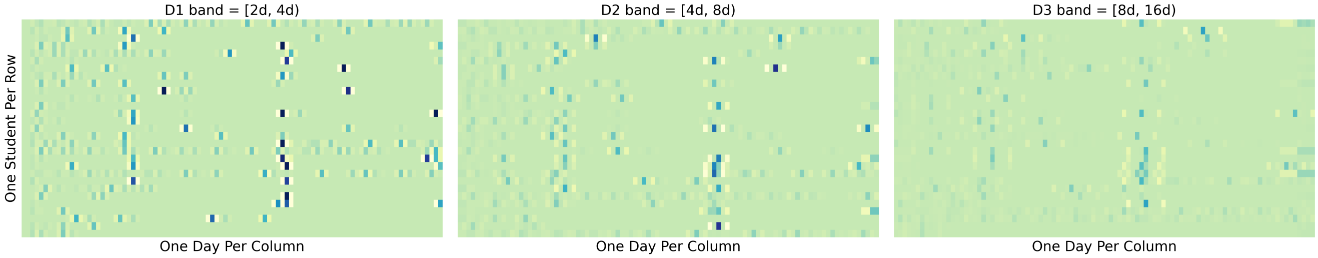

Figure 2: The SWT Signals in the SP18 Dataset. From left to right, the SWT signals D1, D2, D3 capture practice sessions in the period bands [2d, 4d], [4d, 8d], and [8d, 16], small-, medium- and largely-spaced respectively. These signals capture practice sessions spaced by different periods. From D1 (more detail) to D3 (more context), we can see the focus gradually spreads out when the level increases. For readability, the plot excludes practice sequences not having any change points in the SWT signals.

we set L = 3 in our experiment. Once the SWT algorithm is built, we applied the change point detection algorithm with the penalty = 0.5 to search for sudden changes in the SWT signals [8]. We decided on this penalty value by maximizing the group difference (Section 5.2) and the goodness-of-fit of regression (Section 6.2). The experiment program and data are available at the link for future work[1].

5.1 Characteristics of SWT Signals

The SWT algorithm decomposes the input signal by multi-level filtering. Filtering at a level k extracts information at the frequency band $[1/2^k f, 1/2^{k+1} f]$ where f is the sampling rate of the input. In our datasets, because the sampling rate is 1 sample per day (1 cycle per day), the three decomposition levels (Dk where k = 1,2,3) filter the input in the frequency bands $[1/2, 1/4]$, $[1/4, 1/8]$, and $[1/8, 1/16]$. In other words, the algorithm filters the input into the period (the duration of time of one cycle) bands D1=[2(d)ays, 4d], D2=[4d, 8d], and D3=[8d, 16d]. We map these three bands to small-spaced, medium-spaced, and largely-spaced practice patterns, respectively. Following this interpretation, we expect the SWT signals to identify students' practice sessions spaced by different periods. For example, D1 can identify sessions spaced by 2 to 4 days, which are small-spaced practice.

To further illustrate this characteristic, Figure 2 demonstrates what the algorithm found in the SP18 dataset. The visualization shows the SWT signals at the three levels. We can see that D1 highlights small-spaced practice sessions. The D2 and D3 signals spread their focus and "blur" the sequences not fitting their period bands. Note, there may be redundancy in the information captured by different components. For example, an input sequence having meaningful change points in D3 can also have ones in D1. Overall, the information about practice sessions at different levels provides an insight into how the students distribute their practice over time. In our analysis of distributed practice patterns, we use the number of change points as the *feature* to represent the information from the three SWT signals.

For readability, we use the lower bound of the frequency band to denote the spaced practice patterns. We call the practice patterns found in the D1, D2, D3 signals *2SP* (2-day spaced practice), *4SP*, and *8SP* patterns, respectively. For reference, the input daily practice sessions are called 1SP. In the SP18 dataset, the means (M) / standard deviation (SD) from the three levels are 2SP = 1.83/2.68, 4SP = 1.79/2.63, and 8SP = 1.75/2.63. In the FA18 dataset, the values are 2SP = 1.12/2.29, 4SP =1.27/2.14, and 8SP = 1.37/2.30.

5.2 Marginal Relationship of Spaced Patterns with Exam Grades

We analyzed the relationships between the practice patterns and student grades by the marginal distribution. The Kruskal-Wallies H-test was applied to test if the groups had the same population median (Figure 3). The method was selected because the sample size was small, and therefore the sample might not follow the normal distribution. The results showed that there was only 2SP that appeared to be significant for both datasets (SP18: H=8.89, p=0.01; FA18: H=7.95, p=0.02). The visualization of the distribution showed that in SP18 A students had a higher 2SP (M = 3.12, SD = 3.11) than C (M = 1.50, SD = 2.32) and B (M = 0.72, SD = 1.79); in FA18, the B students had a higher value (M = 2.17, SD = 3.05) than A (M = 0.62, SD = 1.50) and C (M = 0.41, SD = 1.14).

There are more spaced patterns discovered for B students in FA18 but not A students, which suggests there could be other factors in the correlation of their practice with even higher exam grades. For example, engineering and non-engineering students may have/need different practice strategies adapted to their learning conditions. Despite this slight difference across the two semesters, if we focus on the difference between the higher-performing students (A/B) and the C/D/F ones, the result consistently suggests a positive correlation between exam grades and small-spaced prac-

[1] https://github.com/rickchung/edm21-msa

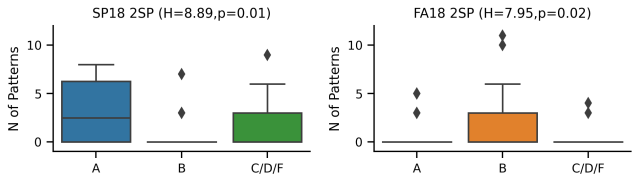

Figure 3: Marginal Distribution of the SP patterns from the Three Grade Groups. The Kruskal-Wallies H-test found that only 2SP was significant in both datasets. The result consistently showed that higher-performing students (A in SP18 or B in FA18) had more small-spaced practice than the C/D/F ones in SP18 and FA18.

tice.

5.3 Quantifying the Schedule of Distributed Practice

We have seen how the SP patterns can help identify practice spaced by different periods. However, this feature alone does not depict the entire picture of distributed practice strategies. Another key factor in the distributed practice effect is the timing of practice. We develop an index to quantify the skewness in practice schedules and investigate its correlation to exam grades. One simple measure of the skewness is the *lag time*. We can use the *lag time* between the occasion of a practice session and a specific event of interest (e.g., exam dates, assignment deadlines) to model the schedule skewness. Due to programming is inherently accumulative, a later exam covers the content from all the previous exams, we cannot assert that a practice session only affects the upcoming exam. Considering this case, we focus on the time lag since the beginning of the semester. Specifically, for an SWT signal at level i, D_i, we can compute the lag of days between the start of the semester and the occurrences of change points. Then, we can transform a practice sequence into a sequence of lags $\{T_1^{D_i}, T_2^{D_i}, T_3^{D_i}, ..., T_n^{D_i}\}$. To know where on the timeline the student has more practice, we compute the sample mean, $\mu_T^{D_i}$. The number, therefore, represents how far the schedule is away from the beginning of the semester. We further divide the number by the total number of days (N_{day}) in the semester for interpretation. The equation of the schedule skewness is defined as

$$SS = \frac{\mu_T^{D_i}}{N_{days}} \tag{4}$$

When a student has all his/her practice sessions early in the semester, SS will be close to zero. If s/he has more practice sessions over the middle of the semester, SS will be some value over 0.5. We can apply the formula to the input signal (1SS) and the SWT signals (2SS, 4SS, 8SS). The result will indicate the schedule of different spaced-practice patterns.

6. MIXED PRACTICE EFFECTS IN MULTIVARIATE ANALYSIS

A distributed practice strategy is multifaceted. The univariate analysis is insufficient because it does not consider the confounding variables. There are two cases remain unclear.

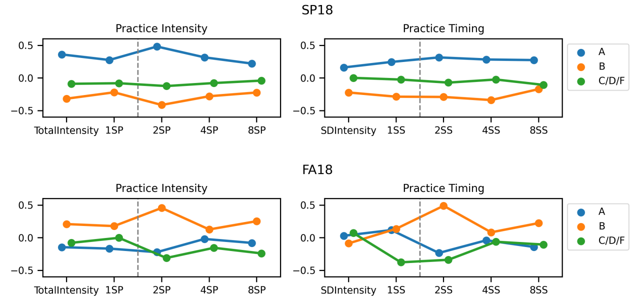

Figure 4: The Standardized Values (Means) and the Interactions of the Basic and Experiment Features. The plot groups the features into intensity and timing according to their functions. The y-axis shows the standardized values (by $(X - M_X)/SD_X$ in each feature category X) for between-features comparison. The vertical dashed lines separate the basic and experiment features. The A students in SP18 have the highest TotalIntensity. On the contrary, the B students in FA18 have the highest TotalIntensity. Although the basic features somehow correlate with the experiment features, we can find more discriminating differences in the experiment spaced-practice patterns.

First, the C/D/F students from SP18 do not have better performance, even though they put efforts into practice just like the better performing students do (A students). Second, in FA18, the analysis does not explain the practice strategy of the A students. They achieve a good grade but do not show significantly more SP. These cases suggest that there could be other factors in the distributed practice effects.

Following this idea, we try to use multivariate analysis that includes experiment SWT features and commonly-used basic features. The basic set comprises the following. For the practice intensity, we use the total number of questions solved (TotalIntensity) and the total number of daily practice sessions (1SP). For the practice schedule, we use the standard deviation of the daily intensity (SDIntensity) and the SS of daily practice sessions (1SS). For reference, we plot the standardized values of the features in Figure 4. The figure shows that although the basic features correlate with the experiment features, we can potentially find more discriminating difference in the spaced practice patterns between the grade groups.

6.1 Assessing the Marginal Effects by Multinominal Logit Regression

To understand the relationship between multiple feature and exam grades, we use the multinominal logistic regression and investigate the marginal effects of the feature values. The *multinominal logistic regression* (MLogit) is a generalized version of the logistic regression for multiclass classification problems [15]. We can use MLogit when the dependent variable in a query is nominal (categorical) and has more than two possible categories. The setup of MLogit is similar to the logistic regression. We assume a linear relationship between the independent variables (predictors), X, and the dependent variable (response), Y, and model the probability of the $Y \in \{y_1, ..., y_k\}$ by the logistic function (*sigmoid*) and k-1 sets of weights (w_k for the label y_k):

$$P(Y = y_k|X = X_1, ..., X_n) = \frac{exp(w_{k0} + \sum_i w_{ki}X_i)}{1 + \sum_j exp(w_{j0} + \sum_i w_{ji}X_i)} \tag{5}$$

We can obtain the prediction by picking up the class with the highest probability. The main advantage of MLogit over other classification techniques is the interpretability. We can explain the contribution of individual features to the output probability (dx/dy) similar to the linear regression [15]. In the analysis, we set Y as the grade groups (A, B, C/D/E) and examine the marginal effects with respect to X when the model fits different sets of predictors.

6.2 Comparing Alternative Models

To understand the capability and limitation of the SWT model of distributed practice, we use MLogit to fit various baseline and experiment feature sets. Afterward, we benchmark the quality of these models by the goodness-of-fit. Due to MLogit does not use the standard R^2, we use the measure of the goodness-of-fit by *McFadden's pseudo* R^2 [10]. McFadden's pseudo R^2 uses the formula

$$R^2 = 1 - \frac{\ln L(M_{full})}{\ln L(M_{intercept})} \tag{6}$$

where L is the estimated likelihood. A small ratio of the two log-likelihoods (or a large McFadden's pseudo R^2) suggests that the full model is better than the intercept model. We can use this measure to benchmark one model against another if they fit the same data.

We compared the experiment and alternative baseline models. The result showed that none of the baseline models were competitive with even the simplest SWT model (using only 2SP, 4SP, 8SP). The best baseline model ($M_{BaseAll}$) used all the baseline variables and achieved R^2 = 0.04 in SP18 and R^2 = 0.07 in FA18. The simplest SWT model ($M_{ExpDose}$) achieved R^2 = 0.07 in SP18 and R^2 = 0.09 in FA18. The best experiment model (M_{ExpAll}) used all the SWT variables and achieved R^2 = 0.12 in both SP18 and FA18. Using all the baseline and experiment variables, the ensemble model ($M_{Ensemble}$) unsurprisingly outperformed all the other models and achieved R^2 = 0.13 and R^2 = 0.23 in SP18 and FA18, respectively.

6.3 Marginal Effects in the Regression Models

In SP18, $M_{ExpDose}$ found 2SP was a significantly-positive predictor for the A students (dx/dy = 0.07, p = 0.01). M_{ExpAll} also found that 2SP was a significant predictor for the A students (dx/dy = 0.10, p = 0.00). Besides, it found 4SS and 8SS were significant for the C/D/F students (dx/dy = 2.55, p = 0.02; dx/dy = -2.58, p = 0.03). In FA18, $M_{ExpDose}$ found 2SP was significantly-positive predictor for the B students (dx/dy = 0.12, p = 0.00). M_{ExpAll}, however, did not find any significant predictor. Part of the result is similar to the analysis of marginal distribution. In SP18, an increase of small-spaced practice adds to the likelihood of A. In FA18, the same effect works for B. It is worth noting an additional finding in $M_{Ensemble}$ from SP18. When we control the intensity and SS, the model shows two extra significant predictors for the grade C/D/F: 4SS and 8SS. The marginal effect suggests that an increase/decrease in 4SS/8SS adds to/reduces the likelihood of C. Since an increase in SS means the schedule becomes later in the semester, these two findings somewhat suggest the same thing: students who practice early and space the practice largely are less likely to obtain C/D/F.

It is also worth noting that the one in FA18 improves the most from the best experiment model and reaches R^2 = 0.23. When predicting the A students, the model shows 1SS (dy/dx = 1.01, p = 0.00) and the total intensity (dy/dx = -0.02, p = 0.04) are significant predictors; when the model predicts the C students, 1SS is the only significantly-negative predictor (dy/dx = -0.90, p = 0.01). We do not find the same effect in any of the baseline models. The result complements a missing part of our analysis about the A and C students' practice strategies in FA18. It suggests that an increase in 1SS adds to the likelihood of A. Conversely, the same increase reduces the one of C/D/F. In other words, more early or late practices in the semester may reduce or improve the probability of C/D/F or A, respectively.

7. CONCLUSIONS

Students' practice behavior is challenging to model because they can practice anytime and do not necessarily follow a unified schedule. This study aims to build such a feature model that can help researchers describe the distributed practice behavior. We adopted the method from multiresolution analysis to extract patterns of our distributed practices, focusing on two factors in the distributed practice effect: intensity and timing. In the experiment, we applied the MRA model and extracted features that could represent practices spaced by different periods, including small (2-4 days), medium (4-8 days), and large (8-16 days). These three kinds of practice patterns were analyzed to explain their correlation to the exam grades. We found that students who practiced early and spaced the practice by the small and large periods were more likely to get a higher grade than C/D/F. Also, the students having more small-spaced practices throughout the semester (i.e., practicing more persistently) were more likely to get better exam grades. Additionally, the MRA model was benchmarked against baseline models. The result showed that the MRA model not only achieved a better goodness-of-fit than the baselines when working alone, but it could complement a baseline model and achieve better performance.

8. REFERENCES

[1] R. R. Coifman and D. L. Donoho. Translation-Invariant De-Noising. *Wavelets and Statistics*, pages 125–150, 1995.

[2] I. Daubechies. *Ten Lectures on Wavelets.* Society for Industrial and Applied Mathematics, 1 1992.

[3] P. F. Delaney, P. P. Verkoeijen, and A. Spirgel. Spacing and Testing Effects. In *Psychology of Learning and Motivation - Advances in Research and Theory*, volume 53, pages 63–147. Elsevier Inc., 1 edition, 2010.

[4] O. Dermy, A. Brun, and U. D. Lorraine. Can we Take Advantage of Time-Interval Pattern Mining to Model Students Activity ? In *Proceedings of the 13th International Conference on Educational Data Mining, EDM 2020*, pages 69–80, 2020.

[5] J. Dunlosky, K. A. Rawson, E. J. Marsh, M. J. Nathan, and D. T. Willingham. Improving Students' Learning With Effective Learning Techniques. *Psychological Science in the Public Interest*, 14(1):4–58, 1 2013.

[6] J. Fowler. The redundant discrete wavelet transform and additive noise. *IEEE Signal Processing Letters*, 12(9):629–632, 9 2005.

[7] N. Gitinabard, T. Barnes, S. Heckman, and C. F. Lynch. What will you do next? A sequence analysis on the student transitions between online platforms in blended courses. In *Proceedings of the 12th International Conference on Educational Data Mining, EDM 2019*, pages 59–68, 2019.

[8] R. Killick, P. Fearnhead, and I. A. Eckley. Optimal Detection of Changepoints With a Linear Computational Cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 12 2012.

[9] G. Lee, R. Gommers, F. Waselewski, K. Wohlfahrt, and A. O'Leary. PyWavelets: A Python package for wavelet analysis. *Journal of Open Source Software*, 4(36):1237, 4 2019.

[10] J. S. Long and J. Freese. *Regression models for categorical dependent variables using Stata*, volume 7. Stata press, 2006.

[11] N. R. Mabroukeh and C. I. Ezeife. A taxonomy of sequential pattern mining algorithms. *ACM Computing Surveys*, 43(1):1–41, 11 2010.

[12] S. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 7 1989.

[13] E. Panadero. A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, 8(APR):1–28, 2017.

[14] D. B. Percival and A. T. Walden. *Wavelet Methods for Time Series Analysis*. Cambridge University Press, Cambridge, 2000.

[15] R. L. Strawderman, A. C. Cameron, and P. K. Trivedi. Regression Analysis of Count Data. *Journal of the American Statistical Association*, 94(447):984, 9 1999.

[16] K. VanLehn. The Behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*, 16(3):227–265, 2006.

[17] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–511–I–518. IEEE Comput. Soc, 2001.

[18] Y. Zhang, Y. Dang, and B. Amer. A Large-Scale Blended and Flipped Class: Class Design and Investigation of Factors Influencing Students' Intention to Learn. *IEEE Transactions on Education*, 59(4):263–273, 11 2016.