2021 ASEE ANNUAL CONFERENCE
Virtual Meeting | July 26–29, 2021 | Pacific Daylight Time
Paper ID #34129

# Quantifying the Impact of Students' Semester Course Load on Their Academic Performance

**Mr. Shahab Boumi, University of Central Florida**

Shahab Boumi is a Ph.D. student in the department of Industrial Engineering and Management Systems (IEMS) at the University of Central Florida. His main research focus on investigating students behavioral hidden patterns in large scale data sets using supervised/unsupervised machine learning methods, stochastic processes, and optimization tools.

**Prof. Adan Ernesto Vela, University of Central Florida**

# Quantifying the impact of students' semester course load on their academic performance


## Abstract

Students' academic success in science, technology, engineering, and mathematics (STEM) careers is one of the most popular subjects that has gained attention among educational researchers for decades. Many studies have shown students' educational outcomes can be affected by academic factors including high school GPA, SAT score test, student admission type (transfer or first-time-in-college), as well as demographic features such as gender, ethnicity, and family income. Additional studies have investigated the relationship between students' course load and their academic outcomes. In this paper, we define students' course load based on the number of courses they take each semester, which is assumed to have a discrete probability distribution. To assess if students' course load impacts their academic performance, we apply Hidden Markov models, an unsupervised learning method, to classify students into three categories: high-level enrollment, medium-level enrollment, and low-level enrollment. The sequence of the number of courses students enroll each semester during their academic career is fed into our proposed model as input. The output which is a qualitative measure and is not directly observable is the estimated enrollment level for the students. After students' classification, we derive and compare their academic (e.g., cumulative GPA, graduation rate, and DFW rate) and non-academic (e.g., family income level) features for each enrollment level category. Findings show that students who have more engagement with the university have higher academic performance (higher cumulative GPA and graduation, lower DFW rate) than those with lower engagement. Our analysis also demonstrates that students from families with low-income levels are more likely to have lower enrollment levels. Such results indicate that university managers can improve students' educational performance and, subsequently, the university graduation rate by encouraging students to engage more with the university by providing academic and financial support. These results are based on the data collected from the University of Central Florida from 2008 to 2016 and contain approximately 170,000 students.


## Introduction

Predicting students' academic success in higher education has been one of the most popular subjects among educational researchers for decades. Identifying vulnerable students early on can help in providing them critical supports through their academic careers. Based on the National Student Clearinghouse Research Center, on average, only 58% of students who started colleges in fall 2012 have earned their degrees within six years, which means 42% of students either have graduated in more than six years or have left college without degrees [1]. Research show that demographic features, such as gender, race, and family income have effects on the six-year graduation rate and academic persistence [2, 3]. For example, the retention rate for Hispanic and African-Americans is lower compared to White and Asian students [4].

Besides the demographic features, other studies have investigated the relationship between course load and students' academic outcomes. Based on their results, there is no evidence that lighter course load causes higher academic success. In other words, most studies have not found a negative correlation between course load and academic performance. For example, Shami et al. [5] showed that in a Saudi university with six department, for a given semester, students who take less than 12 credits have the lowest GPA, while students with more than 17 credits have the highest GPA in the semester. In a similar study, Zakir Khouj et al. [6] investigate the relationship between credit load and GPA for students who enrolled for 13 to 18 semester credits. The results showed that there is a positive relationship between student' credit load and their academic performance. Huntington-Klein and Gill [7] in their research studied the impact of semester course load on students' GPA. They illustrated that there is no evidence that high course load have a negative impact on students' GPA, even for students with low academic performance. Attewell et al. [8] demonstrated that students who start their education with higher credit load are more likely to have a greater level and commitment to their academic goal and consequently have a higher performance compared to the students with lower credit load.

All these studies consider credit load based on students' academic behavior in a *single* semester, ignoring students' *long term* mindsets for academic engagement. As an example, a student might not have a high course load in his/her first semester, while his/her enrollment level over the course of his/her career is high on average. In this study, we define a student's course load in a semester as the number of courses he takes in that semester. Considering this definition, unlike students' course load, which is a quantitative measure and is directly observable for each specific semester, students' overall attitude for academic enrollment through their course of study is a qualitative factor and therefore hidden. We identify this overall attitude as *enrollment level*, which depends on the student course load in his/her current semester and all previous semesters. We use a multi-period dynamic approach to classify students into three enrollment levels: low, medium, and high, and investigate the relationship between students' enrollment levels and students' academic performance, including cumulative GPA, DFW rates, and graduation rate.

The remainder of the paper is organized as follows: the problem is described in the Problem Statement section, Methodology section explains the proposed approach for solving the problem. In the Data Description part, we provide more details about our data set, and Parameters Estimation, covers the parameter estimations for our model. Section Results presents the numerical results and the paper is concluded in conclusion section.

## Problem statement

This paper is aimed at developing a model to distinguish between student's semester course load and student's enrollment levels. We apply the Hidden Markov Model (HMM), which takes a sequence of students' course load (any number from 1 to 7) and estimates students' enrollment levels (low, medium, or high). Four examples of student course load (input) and corresponding student enrollment level (output) are provided in Table 1. Student number 1 has registered six semesters at University of Central Florida, and the number of courses he has taken from semester 1 to semester 6 is 5,5,2,5,6 and 5, respectively. In most of the semesters (except the third semester), he has heavy course load. Our proposed model estimates a high enrollment level for this student each semester. In other words, just one low course load semester (the third semester) does not impact his/her enrollment level in the long term. However, for student number 4, the enrollment level is changed after his/her third semester from high to low.

Table 1: Example for students' semester course load and corresponding enrollment levels

| Student number | course load | Enrollment Level |
|---|---|---|
| 1 | 5,5,2,5,6,5 | H, H, H, H, H, H |
| 2 | 3,3,3,4 | M, M, M, M |
| 3 | 1,2,1,1,2 | L, L, L, L, L |
| 4 | 5,5,5,1,2,1 | H, H, H, L, L, L |

## Methodology

In the literature, different methods are applied to assess students' academic performance and persistence in higher education, among which stochastic methods and machine learning algorithms are the most popular approaches [9, 10, 11, 12, 13]. In this paper, we use Hidden Markov Model (HMM) to classify students based on their engagement levels during their academic careers. Many studies have used HMM to model and investigate students' educational behaviors [14**?** , 15, 16, 17, 18]. For example, Boumi and Vela [19] applied HMM to classify students based on their enrollment strategies and then compute and calculate academic performance for each class of students. Results indicate that students with a full-time enrollment strategy have higher educational outcomes compared to students with mixed and part-time enrollment strategies. In another study, Chen et al. [20] use HMM to categorize students based on their learning styles and investigate the relationship between learning style states and learning efficiency in massive open online courses (MOOCS).

Hidden Markov Model, like any other Markov Models has a set of states $Q = \{q_1, q_2, \ldots, q_N\}$. However, in HMM, the states' status are hidden, and the model aims to estimate the states given a sequence of observations $O = o_1, o_2, \ldots, o_T$ with the length of $T$, where each observation is drawn from the set of $M$ possible observations $V = \{v_1, v_2, \ldots, v_M\}$. Each Hidden Markov Model has three parameters: (1) $\pi_0$, is the initial probability distribution by which the system begins; (2) $A$, the transition matrix which shows the probabilities for the system to move between the hidden states; and (3) $B = b_i(o_t)$, the emission matrix that shows the probability of generating observation $o_t$ when the system is in state $qi$. The first step to develop an HMM is to estimate parameter $\lambda = (\pi_0, A, B)$. We apply the Baum-Welch algorithm that is an expected-maximization method which starts with an initial guess for the parameters. This method then iteratively improves the estimations by calculating the likelihood of any sequence of observations given $\lambda$, until converging to optimal values ($\lambda^*$). After estimating model parameters, we use the Viterbi algorithm to estimate the hidden states. Viterbi algorithm takes a sequence of observations and $\lambda^*$ as inputs and estimates the sequence of hidden states by calculating the likelihood of the observation sequences.

## Data Description

This study is conducted based on data collected from the University of Central Florida from 2008 to 2017. The data set includes a variety of information for approximately 170,000 undergraduate students, including demographic (e.g., gender, race, age, and reported family income), students' admission type (first-time-in-college or transfer), students' academic load (full-time or part-time), students' academic level (freshman, sophomore, junior, and senior), courses taken by the students in each semester with corresponding grades, degree awarded, and so on. Some of this information are summarized in Table 2 through Table 4. These tables show that UCF is a unique university in some aspects including the large population of Hispanic students and transfer students.

In order to estimate each student's enrollment level, we collect the number of courses that student has taken in each semester as a sequence. Furthermore, since some students do not take courses in summer semesters, we have limited our analysis to the sequence of Fall and Spring semesters.

Table 2: Students' gender distribution at UCF

| | **Male** | **Female** |
|---|---|---|
| Percentage | 44% | 56% |

Table 3: Students' race distribution at UCF

| | **White** | **Hispanic** | **African-Am.** | **Other**[1] |
|---|---|---|---|---|
| Percentage | 55% | 24% | 11% | 10% |

Table 4: Students' admission type distribution at UCF

| | **FTIC** | **Transfer** |
|---|---|---|
| Percentage | 41% | 59% |

## Parameters Estimation

In this section, we apply the Baum-Welch algorithms to estimate model parameters $\lambda = (\pi_0, A, B)$. We collect students' course load sequences (HMM observations) that include the number of courses students take in each semester. In this study, we have assumed 7 different values for the HMM observations (number of courses in each semester). Observations 1 through 6 correspond to the semesters with 1 to 6 courses, and observation 7 is assigned to the semesters in which students take 7 or more classes. Results for estimated parameters are provided below:

$$A = \begin{bmatrix} 0.88 & 0.10 & 0.02 \\ 0.09 & 0.84 & 0.07 \\ 0.03 & 0.15 & 0.82 \end{bmatrix}$$

[1] The other category includes American-Indian, Asian, Native Hawaiian, and Multi-racial ethnicity

$$B = \begin{bmatrix} 0.23 & 0.38 & 0.30 & 0.08 & 0.01 & 0 & 0 \\ 0.01 & 0.01 & 0.07 & 0.54 & 0.33 & 0.04 & 0 \\ 0 & 0 & 0.01 & 0.09 & 0.38 & 0.34 & 0.18 \end{bmatrix}$$

$$\pi_0 = \begin{bmatrix} 0.218 & 0.521 & 0.261 \end{bmatrix} \quad \pi = \begin{bmatrix} 0.370 & 0.424 & 0.206 \end{bmatrix}$$

The transition matrix A shows the probability of students moving between different levels of enrollment in two consecutive semesters. Rows 1, 2, and 3 correspond to the low, medium, and high enrollment levels, respectively (the same order for columns 1 to 3). As this matrix suggests, the chances for students with a low enrollment level in semester t to stay in the same enrollment level for semester t+1 is 88%, and they move to the medium and high enrollment level with 10% and 2% probabilities, respectively. Furthermore, the numbers on the main diagonal of the matrix show that students with any levels of enrollment prefer to keep their enrollment level from one semester to the next.

The emission matrix B presents the probability of taking different numbers of classes given students' enrollment levels. For example, for students with a low enrollment level (the first row), the probability of taking 1, 2, and 3 courses in a semester are 23%, 38%, and 30%, respectively (columns 1 to 3). Based on this matrix, while students with the medium enrollment level (the second row) usually take 4 and 5 courses in a given semester, high enrollment level students (the third row) take more than 5 classes in a given semester with higher than 50% chances (sum of columns 6 and 7).

The initial matrix $\pi_0$ shows the probability distribution by which students begin their academic careers with different enrollment levels. For instance, while almost half of the students (52.1%) at UCF start with medium enrollment level, 21.8% and 26.1% of students begin with low and high enrollment levels, respectively. Matrix $\pi$, demonstrates the stationary probability distribution for the transition matrix A. The elements of $\pi$ show the probabilities of students having different levels of enrollment in any given semester. Comparing $\pi_0$ and $\pi$ explains how students' preference over enrollment levels have been changed during their academic careers. For example, students at UCF prefer to have a low enrollment level in later semesters compared to their first semester (37.0% vs. 21.8%).

## Results

In the previous section, we estimated HMM model parameters using the Baum-Welch algorithm. In this section, we use the estimated parameters alongside the Viterbi algorithm to estimate students' enrollment levels. Students are classified into three categories based on their estimated enrollment levels: low, medium, and high enrollment levels. After students' classification, we derive and compare their non-academic (e.g., family income level, gender, and race) and academic (e.g., cumulative GPA, graduation rate, and DFW rate) features for each enrollment level category.

Figure 1 shows the probability distribution over students' enrollment level at UCF. Low, medium, and high enrollment levels correspond to students who keep a constant enrollment level during their academic careers. The fourth group, which is named as *Other*, consists of students who employ a combination of enrollment levels during their education. Student number 4 in Table 1 is

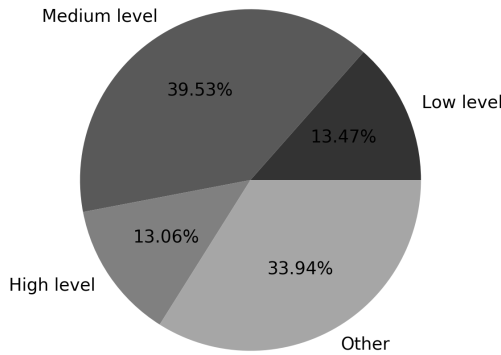


Figure 1: Distribution over enrollment levels at UCF

Table 5: Gender distribution for students with different enrollment levels at UCF

| | **Low** | **Medium** | **High** | **Other** |
|---|---|---|---|---|
| Female | 9.2% | 40.5% | 15.3% | 35.0% |
| Male | 13.6% | 41.4% | 11.9% | 33.1% |

an example from this class. Based on the figure, while 33.9% (the *Other* group) of students switch their enrollment level at some semesters, 66.1% (13.5% + 39.5% + 13.1%) of students have a consistent enrollment level. Among students who keep their enrollment levels during their academic careers, the medium enrollment level group has the largest population ratio (39.5%); and these ratios for high (13.1%) and low (13.4%) enrollment level groups are close together.

Table 5 and Table 6 explain the impact of students' gender and race on their enrollment levels. Based on the Table 5, female students are more likely to have a high enrollment level (15.3%) compared to male students (11.9%). On the other hand, the chance of having a low enrollment level is higher for male students (13.6%) as opposed to female students (9.2%) [2]. Similar results are provided in Table 6, suggesting that white students are more likely to have a high enrollment level (14.9%) compared to black students (10.0%) [3]. Considering these results, gender and race are two significant factors affecting the students' enrollment level at UCF.

Average family income for students with different enrollment levels is shown in Figure 2. Based on the figure, students in the low enrollment level group come from families with lower income (43667 $) compared to the other groups (75529 $ and 62344 $ for high and medium enrollment levels, respectively). Table 7 shows the results for Welch's *t-test* effect size with the corresponding p-values, which suggests that the differences in average family income between students with different enrollment levels are statistically significant. These results are intuitive since registering for more classes in a semester corresponds to higher tuition to be paid to the

[2] Conducted chi-squared test shows that male and female students have a different distribution over enrollment levels and the difference is statistically significant (*p-value*=3.9e-12).

[3] Conducted chi-squared test shows that students with different races have a different distribution over enrollment levels and the difference is statistically significant (*p-value*=0).

Table 6: Race distribution for students with different enrollment levels at UCF

| **Ethnicity** | **Low** | **Medium** | **High** | **Other** |
|---|---|---|---|---|
| White | 8.9% | 42.2% | 14.9% | 34.0% |
| Hispanic | 11.2% | 43.6% | 12.9% | 32.3% |
| Black | 12.2% | 47.7% | 10.0% | 30.1% |
| Other race | 8.4% | 39.3% | 17.4% | 34.9% |

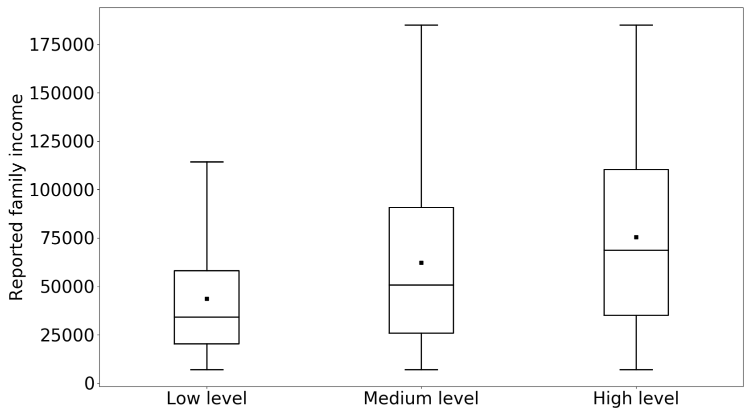


Figure 2: Average family income for students with different enrollment levels

university. These findings can prove promising as it shows vulnerable students can be identified early on during their academic careers and therefore provided financial supports to boost their engagement with the university.

Figure 3 compares the average cumulative GPA for students with different enrollment levels. Based on the figure, students with high enrollment levels have a higher cumulative GPA (3.12) than students with medium (2.92) and low (2.76) enrollment levels. Welch's *t-test* effect size is conducted to assess if the difference in the mean GPA between clusters is significant. Results are summarized in Table 8, which suggests that there is a meaningful difference in the mean cumulative GPA between different enrollment levels. Also, the effect size column indicates that the more is the difference between the enrollment levels, the larger is the computed effect size.

Figure 4 shows the DFW rate computed for students with different enrollment levels. In this paper, the DFW rate for a student is defined as the number of courses with D, F, and W grades divided by the total number of courses the student takes during his academic career. As illustrated in the figure, the low enrollment level group has the highest DFW rate (0.230) among all the

Table 7: Hypothesise results for family income comparison

| **Levels** | **Effect size** | **P-value** |
|---|---|---|
| High vs. Medium | Small effect | 0.0 |
| High vs. Low | Large effect | 0.0 |
| Medium vs. Low | Medium effect | 0.0 |

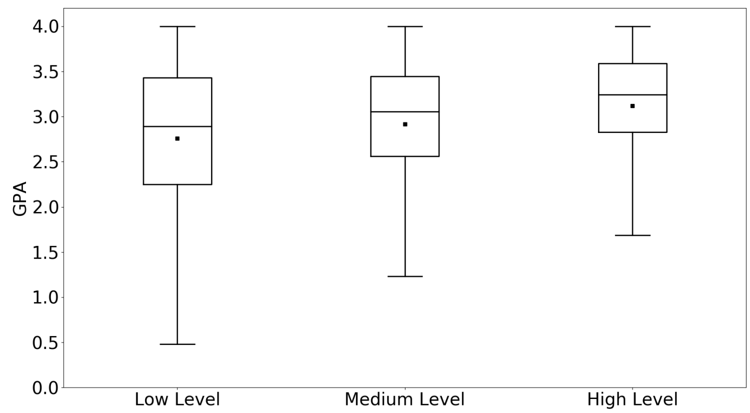


Figure 3: Average cumulative GPA for students with different enrollment levels

Table 8: Hypothesise results for GPA comparison

| **Levels** | **Effect size** | **P-value** |
|---|---|---|
| High vs. Medium | Small effect | 1.24e-302 |
| High vs. Low | Medium effect | 0 |
| Medium vs. Low | Small effect | 4.47e-102 |

groups, followed by the medium (0.155) and high (0.105) enrollment levels, respectively. The results for the effect size tests in Table 9 show that the differences in the DFW rates between these three groups are statistically significant.

Finally, we compare academic performance of students with different enrollment levels in terms of graduation rate. American universities use the six-year graduation rate method to evaluate students' performance. In higher education system, different methods are used to estimate graduation rate [21]. Federal regulations define the six-year graduation rate as the number of FTIC students who earn their degrees within the 150% of the standard time for degree completion [22]. Therefore for a four-year degree program, only students who finish the program within six

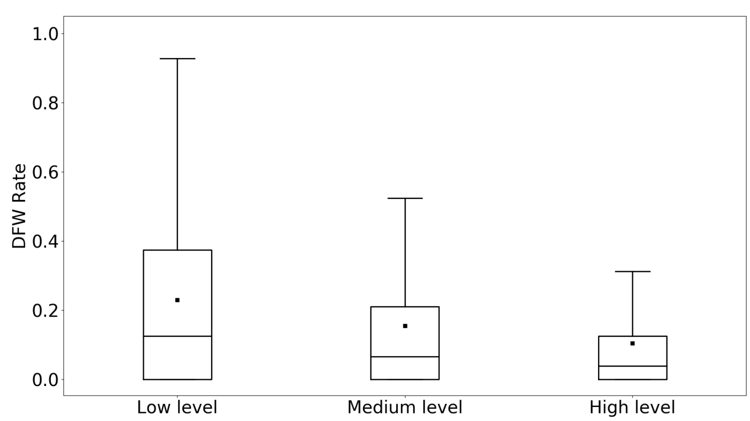


Figure 4: Average DFW rates for students with different enrollment levels

Table 9: Hypothesise results for DFW rate comparison

| Levels | Effect size | P-value |
|---|---|---|
| High vs. Medium | Small effect | 6.8e-215 |
| High vs. Low | Medium effect | 0.0 |
| Medium vs. Low | Small effect | 6.7e-170 |

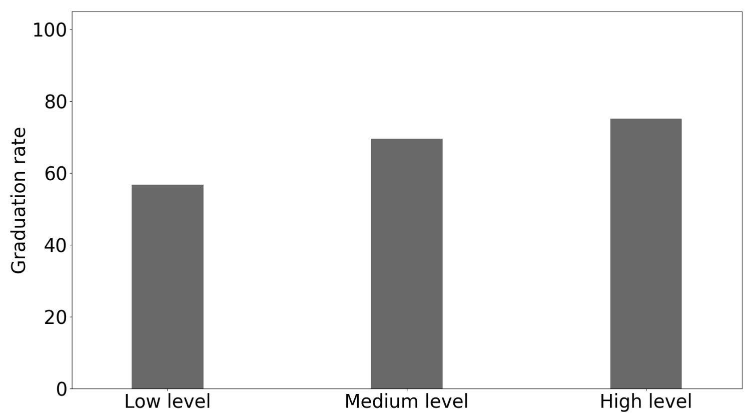


Figure 5: Graduation rates for students with different enrollment levels

years are considered graduates. Since students with a low enrollment level, on average, take fewer courses than students with medium and high enrollment levels, they are more likely to earn a degree in more than six years. The six-year graduation rate method assumes such students leave school with no degrees and, therefore, might underestimate the graduation rate. In this paper, we compute the graduation rate for students with different enrollment levels regardless of their time to graduation. Results illustrated inFigure 5 indicate that students with a low enrollment level have the lowest graduation rate (56.8%), followed by medium (69.5%) and high (75.2%) enrollment levels.

## Conclusion

The main contribution of this study is to distinguish between students' semester course load and students' enrollment levels. We define students' semester course load as the number of credits students take each semester. However, students' enrollment levels are considered the students' attitudes in taking courses during their academic careers. Unlike the student's course load that is a quantitative measure and is directly observable for each student, student's enrollment levels is a qualitative measure and is hidden. We apply the Hidden Markov Model (HMM) to estimate students' enrollment levels using the sequences of their course load in each semester. We then classify them into three categories based on their enrollment levels, including low, medium, and high. Comparing students' academic performance between these categories shows that students with high enrollment levels have the best educational outcome among other categories, followed by medium and low enrollment levels. Furthermore, our financial analysis demonstrates that family income is less for students with a low enrollment level compared to the other two categories. Such results can help university policymakers in the early identification of more vulnerable students and providing them with financial and academic supports to boost their engagement with the university.

## References


[1] Doug Shapiro, Afet Dundar, Faye Huie, Phoebe Khasiala Wakhungu, Ayesha Bhimdiwala, and Sean Eric Wilson. Completing college: A national view of student completion rates–fall 2012 cohort (signature report no. 16). *National Student Clearinghouse*, 2018.

[2] Leila Zahedi, Stephanie J Lunn, S Pouyanfar, MS Ross, and MW Ohland. Leveraging machine learning techniques to analyze computing persistence in undergraduate programs. In *2020 ASEE Virtual Annual Conference Content Access*, number 10.18260/1, pages 2–34921, 2020.

[3] Stephanie Lunn, Leila Zahedi, Monique Ross, and Matthew Ohland. Exploration of intersectionality and computer science demographics: Understanding the historical context of shifts in participation. *ACM Transactions on Computing Education (TOCE)*, 21(2):1–30, 2021.

[4] William G Spady. Dropouts from higher education: Toward an empirical model. *Interchange*, 2(3):38–62, 1971.

[5] Ahmed Shami, Mohammad Ansar, et al. Relationship of students' course loads with their grade point average scores. 1980.

[6] Abdullah Mohammad Zakir Khouj, Ahmed Shami, and Mohammad Ansar. Relationship of students' course loads with their grade point average scores, study ii: An occasional paper. 1982.

[7] Nick Huntington-Klein and Andrew Gill. Semester course load and student performance. *Research in Higher Education*, pages 1–28, 2020.

[8] Paul Attewell, Scott Heil, and Liza Reisel. What is academic momentum? and does it matter? *Educational Evaluation and Policy Analysis*, 34(1):27–44, 2012.

[9] Mehrdad Mirzaei and Shaghayegh Sahebi. Modeling students' behavior using sequential patterns to predict their performance. In *International Conference on Artificial Intelligence in Education*, pages 350–353. Springer, 2019.

[10] Hossein EbrahimiNejad, Hassan Ali Al Yagoub, George D Ricco, Matthew W Ohland, and Leila Zahedi. Pathways and outcomes of rural students in engineering. In *2019 IEEE Frontiers in Education Conference (FIE)*, pages 1–6. IEEE, 2019.

[11] RR Kabra and RS Bichkar. Performance prediction of engineering students using decision trees. *International Journal of computer applications*, 36(11):8–12, 2011.

[12] Mehrdad Mirzaei. *Discriminative Factorization Models for Student Behavioral Pattern Detection and Classification*. PhD thesis, State University of New York at Albany, 2020.

[13] Zhila Aghajari, Deniz Sonmez Unal, Mesut Erhan Unal, Ligia Gómez, and Erin Walker. Decomposition of response time to give better prediction of children's reading comprehension. *International Educational Data Mining Society*, 2020.

[14] Manie Tadayon and Gregory J Pottie. Predicting student performance in an educational game using a hidden markov model. *IEEE Transactions on Education*, 2020.

[15] Kristy Elizabeth Boyer, Robert Phillips, Amy Ingram, Eun Young Ha, Michael Wallis, Mladen Vouk, and James Lester. Investigating the relationship between dialogue structure and tutoring effectiveness: a hidden markov modeling approach. *International Journal of Artificial Intelligence in Education*, 21(1-2):65–81, 2011.

[16] Shahab Boumi, Adan Vela, and Jacquelyn Chini. Quantifying the relationship between student enrollment patterns and student performance. *arXiv preprint arXiv:2003.10874*, 2020.

[17] Mohammad H Falakmasir, José P González-Brenes, Geoffrey J Gordon, and Kristen E DiCerbo. A data-driven approach for inferring student proficiency from game activity logs. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*, pages 341–349. ACM, 2016.

[18] Loc Nguyen. A new approach for modeling and discovering learning styles by using hidden markov model. *Global Journal of Human Social Science Linguistics & Education*, 13(4), 2013.

[19] Shahab Boumi and Adan Vela. Application of hidden markov models to quantify the impact of enrollment patterns on student performance. *International Educational Data Mining Society*, 2019.

[20] Haijian Chen, Yonghui Dai, Heyu Gao, Dongmei Han, and Shan Li. Classification and analysis of moocs learner's state: The study of hidden markov model. *Computer Science and Information Systems*, 16(3):849–865, 2019.

[21] Shahab Boumi and Adan Ernesto Vela. Improving graduation rate estimates using regularly updating multi-level absorbing markov chains. *Education Sciences*, 10(12):377, 2020.

[22] Linda Serra Hagedorn. How to define retention. *College student retention formula for student success*, pages 90–105, 2005.