A click is more than a click: Exploring the relation between students’ online learning behavior and course performance by increasing the level of contextual information in data analysis.

Zhongzhou Chen*, Mengyu Xu, Geoffrey Garrido, and Matthew W. Guthrie
Department of Physics, University of Central Florida, 4111 Libra Drive, Orlando, Florida, 32816
Department of Statistics and Data Science, University of Central Florida, 4000 Central Florida Blvd., Orlando, Florida, 32816

This study examines whether including more contextual information in data analysis could improve our ability to identify the relation between students’ online learning behavior and overall performance in an introductory physics course. We created four linear regression models correlating students’ pass-fail events in a sequence of online learning modules with their normalized total course score. Each model takes into account an additional level of contextual information than the previous one, such as student learning strategy and duration of assessment attempts. Each of the latter three models is also accompanied by a visual representation of students’ interaction states on each learning module. We found that the best performing model is the one that includes the most contextual information, including instruction condition, internal condition, and learning strategy. The model shows that while most students failed on the most challenging learning module, those with normal learning behavior are more likely to obtain higher total course, whereas students who resorted to guessing on the assessments of subsequent modules tended to receive lower total scores. Our result suggests that considering more contextual information related to each event can be an effective method to improve the quality of learning analytics, leading to more accurate and actionable recommendations for instructors.

I. Introduction

Online learning platforms provide a rich variety of data on students’ learning behavior, enabling researchers to explore the relation between learning behavior and learning outcome, motivation, course completion, and other student characteristics. For example, Kortemeyer [1,2] examined both the relation between frequency of material access and students’ course outcome, and the relation between discussion forum posts and learning outcome; Form anek et. al. [3] studied the relation between number of video views, discussion forum participation, peer grading participation and students’ level of motivation and engagement in a massive open online course (MOOC); Lin et. al [4] correlated students’ access of instructional videos with course performance. In the broader field of learning analytics, more sophisticated analytic methods and algorithms have been developed to either identify patterns in students’ online learning behavior [5–7] or predict academic achievement based on large data sets [8–12].

In most of those studies, students’ online learning behavior is characterized by the count, frequency or total duration of one or more types of online learning events, such as the number of discussion forum posts or frequency of video views. However, the same type of learning event occurring under different contexts could be generated by distinct types of student learning behavior. For example, a failed problem solving attempt followed by one or more video access or page access events suggests that the student is trying to learn how to solve the problem, while a sequence of failed homework attempts without accessing relevant instructional materials could imply that the student is randomly guessing, especially when the duration of the attempts are short. However, both kinds of failed attempts would contribute equally to the count or frequency of problem attempt data. Gašević et. al. [13] suggested three types of contextual conditions that can have significant impact on learning analytics, based on Winne and Hadwin's self-regulated learning model [14]: Instruction condition: such as the course mode, course content, choice of technology, and instructional design. Internal Condition: such as the level of utilization of learning tools and the learner's level of prior knowledge. Learning products and strategy: including learner's strategy for completing learning tasks, and the quality of the learning product such as annotations or discussion posts.

A number of recent studies have emphasized to varying degrees the context in which online learning events took place, in addition to the number or frequency of events. For example, Wilcox and Pollock [15] examined the impact of four types of contextual information associated
with students’ answering of online conceptual assessments; Seaton et. al. [16] looked at the impact of the time duration of resource access, Alexandron et. al. and Pallazo et. al [17,18] utilized time duration and IP address to detect possible copying behavior in students’ problem-solving events. Seaton et. al. [19] examined the difference in resource usage that took place when students are completing different tasks in a MOOC.

Can outcomes of learning analytics be improved by considering more contextual information associated with each learning event, without increasing the complexity of the analytic methods? In this study, we increased the contextual information associated with each event in three steps, and demonstrated that each step led to increasingly informative descriptions of students' online learning behavior, and more accurate answers to our research question.

The research question that we try to answer in the current study is how students’ online learning behavior relates to their overall performance in a physics course. In other words, do students that are often referred to as “struggling” in a physics course study online learning resources differently from those who perform well in the course, and if so, what are the most characteristic differences?

To answer this question, we collected students' online learning data from a sequence of 10 online learning modules (OLMs) which were assigned as homework to be completed over two weeks. Each module contains an instructional component and an assessment component with 1-2 problems. Previous studies have shown that the mastery-based learning design of the OLMs can not only improve student learning outcome [20,21] but also increase the interpretability and information richness of learning data [22].

The main events analyzed in the current study are the outcomes of each module, as measured by passing, failing, or aborting the assessment component, resulting in 10 events per student. For each pass-fail event, we extracted three types of contextual information: where, when, and how?

More specifically:

1. **Where was it**: on which of the 10 modules did each pass-fail event take place?
2. **When did it happen**: did the pass-fail event take place before or after the student accessed the instructional material in each module, and after how many attempts did the student choose to access the instructional material?

3. **How did it happen**: For each pass-fail event, how much time was spent on solving the problems? Multiple previous studies have linked abnormally short problem-solving duration with either random guessing or answer copying [18,23–27].

Each context corresponds to one of the conditions proposed by Gašević [13]: the “where” reflects the instructional condition of online materials being organized in a sequence of OLMs, the “when” reflects students’ internal state of choosing whether to access the learning resources, and the “how” serves as one indication of the strategy of producing the learning product.

We refer to the combination of a pass-fail event and its associated contextual information as an “interaction state,” or “state” for short. We created three different levels of interaction states with each level including additional contextual information than the previous level, as explained in detail in section III.B. Therefore, each level contains more states than the previous one. Students’ overall performance in the course is measured by their normalized total course score, which includes scores from homework, exams, lab reports, and classroom clicker questions, and directly determines students’ letter grade.

Three linear regression models were constructed to associate each of the three levels of interaction states with students' final course score. To address the issue of multicollinearity [28] between the large number of variables, we selected for each model a subset of significant variables using a regularized linear regression algorithm LASSO [29], and reconstructed the linear models using those LASSO-selected subsets. Complementary to the linear models, we also plotted students’ transition between different states on neighboring modules using a series of parallel coordinate graphs, which is an updated version of the data visualization scheme developed in an earlier study [30]. As detailed in section IV, a complete description of student learning was obtained by combining the linear model with the corresponding parallel coordinate graph for each level.

In section V, we interpret and compare the outcomes of analysis based on the three levels of interaction states, and discuss the benefit of
including increasing amounts of contextual information on each event. We show that, in this case, the inclusion of more contextual information results in more informative descriptions of students’ learning behavior. The model that includes all three types of contextual information reveals a characteristic difference in the way top and bottom students complete certain OLMs, which provides the most accurate actionable recommendations for instructors. We also discuss the implications of the current results for both instructors and education researchers, as well as caveats and future directions of the current study.

II. Study Setup

A. Design of OLM sequence

The OLM sequence is created using the Obojobo learning objects platform developed by the Center for Distributed Learning at University of Central Florida [31]. Each OLM consists of an instructional component (IC) and an assessment component (AC) (c.f. Figure 1). The AC contains 1-2 multiple choice problems and allows a total of 5 attempts. Each of the first 4 attempts are sets of isomorphic problems assessing the same physics knowledge but with different surface features or different numbers. On the 5th attempt, the same problem on the 1st attempt is presented to the students again. On four of the modules used in the current study, a new set of isomorphic problems is presented to students on each of the first 3 attempts, while the same problems on the 1st and 2nd attempt were repeated on the 4th and 5th attempts. Each IC contains a variety of learning resources, including text, figures, videos, and practice problems, focusing on explaining one or two basic concepts or introducing problem solving skills that are assessed by the problems in the AC. Upon opening a new module, a student must make one attempt at the AC before being allowed access to the IC. Access to the IC is locked again when the student makes a new attempt at the AC and is unlocked after the answers are submitted. Students are required to access the OLM sequence in the order given. In the 2017 implementation, due to platform limitations, students could access the next module once they had submitted their 1st attempt on the current module. However, students were not explicitly informed of that information, and were encouraged to complete the current module by either passing the AC or using up all attempts before moving on to the next one.

B. Implementation of OLM sequence

The OLM sequence used in the current study consists of 10 modules covering the subject of Work and Mechanical Energy, implemented in a large calculus-based college introductory physics course. The problems in the AC are inspired by either common homework problems [32] or research-based assessment instrument [33]. Readers can access a sample OLM sequence following [34].

The OLM sequence was assigned to students as homework. Modules 1-6 were released one week before modules 7-10, and all 10 modules were due 2.5 weeks after the release of the first six modules. Completing all 10 modules was worth 9% of the total course score, and each module was weighted equally. The modules were released concurrently with classroom lectures on the same topic. No other assignments were assigned to the students during the 2.5 week period. A total of 230 students attempted at least one module, and 223 students attempted all 10 modules.

III. Methods

A. Analysis of students’ click-stream data

Students’ click-stream data collected from the Obojobo platform are analyzed using R and the tidyverse package [35,36]. For the current study, we extracted the following types of information from the click-stream data:

AC attempt outcome and duration: An AC attempt is recorded as “pass” if the student answers every question correctly, otherwise it is recorded as “fail.” The duration of each attempt is recorded as the time between the student clicks...
a button to start the attempt, and when the student clicks another button to submit the answers. Study sessions: A study session is defined as all students’ interaction with the IC between two consecutive AC attempts on a given module. The duration of a single study session is the sum of all the events that took place during the session, including viewing page content and attempting practice problems. Since each module allows a maximum of 5 attempts, and require one attempt before allowing access to the IC, a student can have a maximum of 4 study sessions. However, we observed that in 93% of the cases, each student only had one study session on a given module. In only 6% of the cases did a student have a second study session longer than 60 seconds and at least 30% as long as the longest study session in that module. For those 6% of the cases, we only consider the first of the two study sessions, which is usually the longer one. In the remaining 1% of the cases, the second (and 3rd) study session are neglected because they are either shorter than 30% of the longest study session, or last less than 60 seconds. These choices are unlikely to impact the outcome of the current analysis, because we only consider whether a student had a study session, and how many attempts were made before and after the study session, not the duration of each study session.

B. Students’ Interaction States with OLM

1. Defining interaction states with increasing levels of contextual information

   Level I (3 states): The first level of interaction states includes information on whether a student passed or failed the AC of a specific module. We define the following three interaction states for each module:
   
   - **Pass (P):** A student passes the AC within the first 3 attempts. The reason for this choice is that: 1) On four of the modules the AC will provide a different problem only on the first 3 attempts, and will repeat the 1st problem on the 4th attempt 2) Many students do not have the knowledge or skill to pass the AC on their 1st attempt, so they essentially have 2 attempts after studying the IC to be considered as pass, which provides some tolerance for “slips,” such as putting in the wrong number in the calculator.
   
   - **Fail (F):** Students who cannot pass the AC within the first 3 attempts. In other words, either passed on the 4th or 5th attempt or failed on all 5 attempts.
   
   - **Abort (A):** Students who did not pass the module and did not use up all 5 attempts before moving on to the next module.

   Information of the specific module on which each state occurred is added by combining the module

| State Name          | State Label | Definition                           | Example       |
|---------------------|-------------|--------------------------------------|---------------|
| Before Study Pass   | BSP         | Pass with no study session           | \{F, P\}      |
| After Study Pass    | ASP         | Pass with study session before the 3rd attempt | \{F, S, F, P\} |
| After Study Fail    | ASF         | Fail with study session before the 3rd attempt | \{F, S, F, F, P\} |
| Late Study          | LS          | Fail with study session after the 3rd attempt | \{F, F, F, S, F,P\} |
| No Study            | NS          | Fail with no study session           | \{F, F, F, F, F\} |
| Abort               | AB          | Fail and did not use all 5 attempts  | \{F, S, F\}   |
number with the above states when constructing the linear regression model, which is described in detail in section C.1.

Level II (six states): The second level adds information about students’ interaction with the IC on top of the three states in Level I. More specifically, we divided students according to whether they had a study session before passing or failing the AC. Table 1 lists the six states in this level, with examples of common sequences of events belonging to each state, using “S” to represent a study session, and “P” or “F” to represent the outcome of each attempt. The rationale for dividing P and F states according to whether the outcome is achieved before or after the study session is straightforward: students who can pass the module before studying are likely to have stronger incoming knowledge than those who passed after studying. On the other hand, those who studied immediately after the first or second failing attempt likely are more motivated to learn than those who studied after more than 3 failed attempts or did not study at all.

Level III (nine states): The third level adds contextual information on how students attempted each AC by dividing BSP, ASP and ASF states according to the duration of attempts. Different cutoff values have been proposed in several earlier studies to distinguish between an abnormally short attempt and a regular problem solving attempt. In the current analysis, we estimated the cutoff to be 35 seconds, by fitting the attempt duration distribution using scale mixtures of skew-normal distribution models, detailed in the next section. On modules 2 and 6, the cutoffs are adjusted to 17 and 24 seconds respectively on attempts after the study session due to shorter overall attempt durations. We assert that students who spent less than the cutoff times on an AC attempt are unlikely to have put in an authentic effort to solve or even read the problem body.

Therefore, we divide each of the BSP, ASP and ASF states into two new states, based on if the students’ attempts are classified as “Brief” or “Normal” based on their attempt durations. For example, the BSP state is divided into BSPB and BSPN (Before Study Pass-Brief and Before Study Pass-Normal). For BSP and ASP, the attempt duration is taken from the passing attempt which is also the last attempt. On ASF, the duration is taken as the longest of the first 3 attempts. The resulting nine interaction states and the relation between the three levels are listed in Table 2.

Table 2: Correspondence between level I, II and III states

| Level I | Level II | Level III |
|---------|----------|-----------|
| P       | BSP      | BSPN      |
|         |          | BSPB      |
|         | ASP      | ASPN      |
|         |          | ASPB      |
| F       | ASF      | ASFN      |
|         |          | ASFB      |
|         | LS       | LS        |
|         | NS       | NS        |
| A       | AB       | AB        |

2. Determining the duration cutoff between Brief and Normal attempts by scale mixtures of skew-normal distribution models

Previous studies showed the cutoff between “Brief” and “Normal” attempts can be determined by fitting the distribution of the problem-solving duration using multi-component mixture models (e.g. [37,38]), finding the cutoff between the shortest component and the second shortest component as demonstrated in Figure 2. In the current study, we fit the distribution of problem solving duration from students’ 1st AC attempt by scale mixtures of skew-normal distribution models.

Figure 2: Example of a mixture model fit of the duration profile with maximum duration of 350s, with the blue line indicating the 35 s cutoff used in the current study for the majority of modules.
attempts collected from all 10 modules, using scale mixtures of normal or skew-normal distributions, since previous studies have suggested that students’ problem solving duration distribution are likely skewed [24,37,38]. There are two reasons for using duration data from the 1st attempt. First, because students are required to make their 1st AC attempt before studying the IC, they are more likely to make a random guess, resulting in a higher peak in the distribution. Second, on the 1st attempt, students who made a “Normal” attempt must read the problem text carefully, whereas students who made a “Brief” attempt likely did not, leading to a larger difference in duration. On their 2nd and 3rd attempts, students may be able to read the problem text faster on some modules where the problems are more similar to the 1st attempt, resulting in smaller difference in duration. For those modules, the cutoff for 2nd and 3rd attempts are being adjusted (see below). The reason for aggregating the duration data from all 10 modules is based on the assumption that “Brief” attempts should be largely independent of the context of the problem, since the student was not actually solving it. Aggregating the duration data will increase the accuracy for estimating the cutoff. Model fitting is conducted with package mixsmsn [39], and details are presented in the Appendix. Based on the results from model fitting, the cutoff between Brief and Normal attempts is initially set at 35 seconds for all modules. To check if this 35 second uniform cutoff is reasonable for all modules and all attempts, we compared it to the mean of the log duration distribution of attempts made both before and after a study session. We use the mean of log-duration distribution since the distribution is approximately log-normal on many modules. Attempts longer than 7200 seconds are excluded as outliers. For attempts before the study session, the mean log-durations of all modules are between 70 - 200 seconds, much longer than 35 seconds, with harder modules having shorter mean durations. For attempts after the study session, on two modules (2 and 6) the mean log-durations are 35 seconds and 52 seconds respectively; only about half as long as the duration of attempts before study on the same modules. Both modules contain conceptual problems, and the problems presented on the 2nd or 3rd attempt are very similar to the one on the 1st attempt. It is reasonable to assume that students can correctly solve the problem on their 2nd or 3rd attempts by looking at the new diagram and without fully reading the problem body again. Therefore, for those two modules, we treat the shortest 15% of the attempts as “Brief,” and adjust the cutoffs to 17 and 24 seconds respectively for attempts after study. The mean duration of all other modules increased on attempts after study.

C. Modeling Linear Relation between Interaction States and Total Course Score

1. Initial construction of linear models

We construct three linear regression models between students’ interaction states on each module and their final course score for each of the three levels of interaction states, in the form of:

\[
y_i = \beta_0 + \sum_{k=2}^{K} \beta_{1,k}M_{i,1,k} + \sum_{k=2}^{K} \beta_{2,k}M_{i,2,k} + \cdots + \sum_{k=2}^{K} \beta_{10,k}M_{i,10,k} + \varepsilon_i
\]

where for the \(i\)th student, \(y_i\) is the standardized final course score with mean of 0 and standard deviation of 1 (referred to as final course z-score in the rest of the paper), \(\varepsilon_i\) represents the “noise” term that accounts for all other effects not explained by the interaction states on the modules. We assume that \(\varepsilon_i\) are identical and independently normally distributed with mean 0. In the model above, \(M_{i,j,k}\) are dummy variables with \(M_{i,j,k} = 1\) if the \(i\)th student has interaction state \(k\) for module \(j\), and \(M_{i,j,k} = 0\) otherwise, for \(i = 1, 2, \ldots, 10; j = 1, 2, \ldots, 10; k = 1, 2, \ldots, K\). The variables \(M_{i,j,k}\) combine information on the module number with interaction states at each level. Consequently, the model parameter \(\beta_0\) represents the expected final course score for students in a “reference state” for every module, while \(\beta_{j,k}\) measures the difference in the final score by being in state \(k\) in module \(j\) compared to the reference state.
For each of the three levels, the reference state is set to be the first state with $k = 1$. According to the three levels, we study the effects with number of states $K$ to be 3, 6 and, 9 respectively. Specifically, the reference state is listed as follows for each level.

I. Final Course z-score~ 3 states. Reference State: P
II. Final Course z-score~ 6 states. Reference State: BSP
III. Final Course z-score~ 9 states. Reference State: BSPN

In each level, the reference state is selected as the interaction state that is most likely associated with the highest level of content knowledge from an instructor’s point of view. The intercept reflects the predicted final course z-score if all modules are in the reference state. For comparison, we also create a baseline linear regression model between the number of modules a student failed and aborted and their final course score:

$$y_i = \alpha_0 + \alpha_F x_{F,i} + \alpha_A x_{A,i} + \epsilon_i$$

(2)

where $y_i$ is the standardized final course z-score

$$(\hat{\beta}_0, \hat{\beta}) = \arg\min_{(\beta_0, \beta)} \sum_{i=1}^{n} (y_i - \beta_0 - M_i^T \beta)^2 + \lambda \| \beta \|_1$$

(3)

for student $i$ and $x_{F,i}$ and $x_{A,i}$ are the number of modules the student failed or aborted, respectively. The parameter $\alpha_0$ stands for the expected score of students who passed all modules and $\alpha_F$ (and $\alpha_A$, respectively) represents the amount of points decreased in the course final z-score for failing (aborting, respectively) one more module.

2. Addressing Collinearity within regression variables using LASSO

Collinearity and Regularized Regression: To construct the linear model (1), we are estimating $p = 10K - 9$ unknown coefficients, i.e., 21, 51, and 81 respectively for level I, II, and III. The fact that $p$ is nonnegligible to the number of students $n = 207$ can induce significant issue in the regression. In particular, it is likely that the space constructed by the predictors is (nearly) singular, which means some of the covariates are (nearly) linear combinations of others. This issue is known as collinearity and it can consequence in a highly inaccurate and unstable, if not nonexistence, model estimation since the ordinary least square solution of 1 relies on the assumption that the covariate space is nonsingular. In presence of collinearity, the estimated relationship can be spurious and redundant, as the effect of one covariate can be replaced by the combination of others.

In remedy of the collinearity, we employ LASSO (Least Absolute Shrinkage and Selection Operator) estimation [29,40], assuming only a small proportion of the states significantly influence the final course score. The LASSO regression regularizes the estimation by imposing a penalty of model size to the square sum of errors, defined as in the following equation:

where the vector $M_i = (M_{i,j,k}, 1 \leq j \leq 10, 1 \leq k \leq K)^T$ contains the binary state dummy variable for each state in all ten modules, and $\beta$ contains the corresponding coefficients. The tuning parameter $\lambda$ controls the strength of penalty in the model, and hence the sparsity of the estimation. We select $\lambda$ by a 10-fold cross validation with the minimum mean squared error.

LASSO estimation assumes that a small subset of $\beta$ is nonzero and is well-known for its model selection consistency under certain conditions (c.f., [41]). In other words, the estimator 3 is able to select the correct subset of features relevant to the overall course score with high probability. That is, with a large sample size, model 3 selects the relevant models and states and excludes the irrelevant with probability near one. We use for feature selection and then regress the final course z-score against the selected modules and states. Let $\hat{S} = \{j: \beta_j \neq 0\}$ be the index set of significant features selected in 3 and $M_{\hat{S}}$ be the design matrix for the corresponding modules and states. We estimate the corresponding coefficients $\hat{\beta}^*$ from the following regression:

$$y = M_{\hat{S}}^T \hat{\beta}^* + \epsilon,$$

(4)

where $y$ and $\epsilon$ are the vector form of the final course z-score and noise respectively. Note in
equation (4), the irrelevant states are not included in the predictors.

**D. Visualizing students’ transition between interaction states in an OLM sequence**

To visualize how students transition between interaction state from one module to the next, we plot data from the 10 modules on a sequence of nine parallel coordinate diagrams. The two vertical axes on each graph represent the interaction states on two adjacent modules. Each student is represented as a line starting from one interaction state on the left axis and ending on another interaction state on the right axis, as shown in Figure 3, Figure 4, and Figure 5. One or more overlapping lines form a path indicating a transition between two interaction states on two adjacent modules, where a horizontal path means that one or more student remained in the same state on the two modules. The student population is divided equally into top 1/3, middle 1/3 and bottom 1/3 cohorts according to their final course score, with each cohort plotted on its own sequence of graphs. The most populated major paths that add up to half of the population within each cohort are highlighted by a yellow line, with the line widths proportional to the size of the major path. The current visualization scheme has two differences from the version in the earlier study [30]: 1. The ordering of states is now based on the results of the linear model. States that are more frequently correlated with lower course grades are placed lower on the graph, with reference state being placed at the top of the graph. 2. Adjacent paths are no longer clustered into a single path, as it cannot be argued that adjacent states are more similar to each other than distant states.

In addition, variables in the linear model selected by the LASSO estimation algorithm are highlighted by three types of labels on the axis: hollow triangles represent selected variables with $\beta_S^*$ not significantly different from zero, solid squares represent variables with $\beta_S^*$ significantly different from 0 at $\alpha < 0.05$ level, and solid spheres represent with $\beta_S^*$ significantly different from 0 at $\alpha < 0.01$ level. Selected variables with $\beta_S^* > 0$ are represented by Dark Cyan (#1A9F76) labels, and those with $\beta_S^* < 0$ are represented by Pollo Blue (#8DA0CB) labels. Each label is repeated three times on the three graphs for the three cohorts.

**IV. Results**

**A. Baseline Model**

The intercept and coefficients of the baseline regression model (Adjusted $R^2 = 0.18$, $F = 16.21, p < 0.01$) are listed in Table 3. As expected, the average final score of students who passed all modules differs significantly from the average of all students, and the number of both failed and aborted modules are negatively correlated with final course score, with correlation coefficients significantly different from zero at $\alpha < 0.01$ level.

Table 3: Estimated coefficients and p-values for the baseline linear regression model (2).

| State | Coeff. ($\alpha$) | p       |
|-------|-------------------|---------|
| Intercept | 0.74         | 0.00**  |
| F       | -0.15           | 0.00**  |
| A       | -0.32           | 0.00**  |

**B. Level I: Three Interaction States**

For level I (three states on each module), 17 out of 21 variables are selected by the LASSO algorithm, resulting in a linear model of Adjusted

Table 4: Estimated coefficients and p-values of regression (4) for the three-state model (Level I).

| Module | State | Coeff. ($\beta^*$) | p       |
|--------|-------|-------------------|---------|
| NA     | Intercept | 0.80         | 0.00**  |
| m1     | A     | -0.14           | 0.61    |
| m1     | F     | -0.24           | 0.08    |
| m2     | A     | -0.31           | 0.49    |
| m2     | F     | -0.13           | 0.36    |
| m3     | F     | -0.06           | 0.72    |
| m4     | A     | -0.50           | 0.03*   |
| m4     | F     | -0.17           | 0.18    |
| m5     | F     | -0.21           | 0.09    |
| m6     | F     | -0.40           | 0.04*   |
| m7     | A     | -0.48           | 0.01*   |
| m7     | F     | -0.32           | 0.03*   |
| m8     | A     | -0.36           | 0.29    |
| m8     | F     | -0.17           | 0.19    |
| m9     | A     | -0.52           | 0.24    |
| m9     | F     | -0.30           | 0.02*   |
| m10    | A     | -1.02           | 0.23    |
| m10    | F     | 0.25            | 0.04*   |
\[ R^2 = 0.20, F = 4.07, df = 189, p < 0.01. \] The coefficients of the model are shown in Table 4. Most of the variables are negatively correlated with the final score, which is expected since the P state is selected as the reference state for each module. In addition to the intercept, six variables have coefficients that are significantly different from zero, in which five of those are on modules 6-10. This is likely because the difficulty of the modules increases towards the end of the sequence. Surprisingly, the F state on module 10 is positively correlated with the final score, indicating that students with high final scores are more likely to fail on this module.

On the parallel coordinate graph (Figure 3), the three states are ordered as P, F, A, since on all modules (except on m10) the coefficients for both F and A states are negative, with the A states being more negative. Four of the six significant variables correspond to the start or end point of a major path. Of which, m7-A is on the end of a significant path in the bottom cohort only; m7-F is at the junction of two major paths on all three cohorts; m9-F is on 2 major paths in the bottom cohort and one major path in the middle cohort; m10-F is on the end of a major path in the middle cohort only.

C. Level II: Six Interaction States

For level II, 24 out of 51 variables are selected by the LASSO algorithm, resulting in a linear model with adjusted \[ R^2 = 0.33, F = 5.268, df = 182, p < 0.01, \] the coefficients of which are shown in Table 5. In addition to the intercept, 10 variables have coefficients that are significantly different from zero at the \( \alpha = 0.05 \) level, one of which, m1-ASF, is significant at the \( \alpha = 0.01 \) level. Most of the variables are

![Parallel Coordinates Graph for Level I Model](image)

**Figure 3:** Parallel coordinate graphs using Level I (three) states.

| Module | State | Coeff.\((\beta^*)\) | p  |
|--------|-------|---------------------|----|
| NA     | Intercept | 0.75                 | 0.00** |
| m1     | ASP   | -0.26                | 0.04*  |
| m1     | ASF   | -0.51                | 0.00** |
| m3     | ASP   | -0.24                | 0.06   |
| m3     | ASF   | -0.40                | 0.02*  |
| m4     | NS    | -0.55                | 0.10   |
| m4     | AB    | -0.36                | 0.16   |
| m5     | LS    | -0.45                | 0.04*  |
| m5     | NS    | -0.21                | 0.43   |
| m5     | AB    | -1.18                | 0.02*  |
| m6     | ASF   | -0.19                | 0.41   |
| m6     | NS    | -0.51                | 0.11   |
| m7     | ASF   | -0.36                | 0.01*  |
| m7     | LS    | -0.51                | 0.06   |
| m7     | NS    | -0.79                | 0.03*  |
| m7     | AB    | -0.57                | 0.02*  |
| m8     | ASP   | 0.31                 | 0.03*  |
| m8     | LS    | -0.22                | 0.33   |
| m8     | NS    | -0.13                | 0.54   |
| m9     | ASP   | 0.11                 | 0.42   |
| m9     | ASF   | -0.14                | 0.35   |
| m9     | NS    | -0.46                | 0.02*  |
| m10    | ASP   | 0.25                 | 0.06   |
| m10    | NS    | 0.29                 | 0.11   |
| m10    | AB    | -1.25                | 0.11   |
negatively correlated with the final score, except for m8-ASP, m9-ASP, m10-ASP, and m10-NS. The ordering of states on the corresponding parallel coordinate graph (Figure 4) reflects the fact that LS, NS and AB states on multiple modules are significantly negatively correlated with final course score. Of the 10 significant variables, 3 of which: m5-LS, m7-NS, m7-AB are not located on any major path in any of the cohorts, and one variable, m5-AB, is located on a small major path in the bottom third cohort. It is likely that those variables reflect the behavior of a small fraction of students with exceptionally low final course score.

Of the remaining 6 significant variables that are also located on at least one major path, there are several noteworthy observations:

1. Among passing states, ASP and BSP (reference state) are similar in their correlation with final course score, except on m1 and m8. On m1, ASP is significantly correlated with final course score, but is also on a major path in all three cohorts. A possible explanation is that the fraction of students with the highest final scores can pass this module, which introduces the definition of kinetic energy, prior to studying the content. Surprisingly, m8-ASP is positively correlated with final course grade compared to m8-BSP, and is on a major path in both the top and middle cohort, but not in the bottom cohort. This implies that students with high final scores are more likely to pass the module after studying the IC rather than passing on their initial attempt. Both m9-ASP and m10-ASP are also positively correlated with final score, with m10-ASP being marginally significant ($p = 0.06$).

2. For failing states (ASF, LS, and NS), m7-ASF still sits on multiple major paths on all three cohorts, suggesting that only a few top students in the class can pass m7, and the IC of m7 not helping the majority of students. On the other hand, m9-NS is connects three major paths

Table 6: Estimated coefficients and p-values of regression (4) for the nine-state model (Level III).

| Module | State | Coeff.(\(\beta^{*}\)) | p   |
|--------|-------|------------------|-----|
| NA     | Intercept      | 0.52             | 0.00** |
| m1     | ASFN            | -0.37            | 0.01* |
| m3     | ASFB            | -0.63            | 0.01** |
| m3     | ASPB            | -0.70            | 0.12  |
| m4     | NS              | -0.53            | 0.08  |
| m4     | AB              | -0.17            | 0.48  |
| m5     | LS              | -0.45            | 0.02* |
| m5     | AB              | -0.98            | 0.05  |
| m6     | NS              | -0.33            | 0.28  |
| m7     | NS              | -0.45            | 0.16  |
| m8     | ASPN            | 0.09             | 0.55  |
| m8     | ASFB            | -0.55            | 0.11  |
| m8     | LS              | -0.50            | 0.02* |
| m8     | BSPB            | -0.33            | 0.05* |
| m8     | NS              | -0.24            | 0.26  |
| m9     | ASFB            | -1.13            | 0.00** |
| m9     | BSPB            | -0.03            | 0.85  |
| m9     | NS              | -0.21            | 0.26  |
| m10    | ASPN            | 0.19             | 0.16  |
| m10    | BSPB            | -0.65            | 0.00** |
| m10    | AB              | -1.35            | 0.07  |
in the bottom cohort, while almost no student occupied that state.

**D. Level III: Nine Interaction States**

For level III, 20 out of 90 variables are selected by the LASSO algorithm (Table 6), producing a minimum linear model with adjusted $R^2 = 0.39, F = 7.69, df = 186, p < 0.001$. 8 of the 20 variables have correlation coefficients that are significantly different from 0 at the $\alpha = 0.05$ level, two of which are significant at the $\alpha = 0.01$ level. Two variables have positive correlation coefficients, but neither are significant.

For the parallel coordinates graph, we noticed that m9-ASFB and m10-BSPB are the only two variables that are significantly correlated with lower final score at the $\alpha = 0.01$ level, and both ASFB and BSPB states also have significant negative correlations on several other modules. In comparison, BSPN (reference state) is positively correlated with the final score (significant positive intercept), while ASFN on most modules are indistinguishable from BSPN since it is not selected by LASSO on any module except m1. To visually represent this large difference between ASFB/BSPB and ASFN/BSPN, we placed BSPB at the bottom of the graph just above AB, where ASPB is placed next to ASPN since our algorithm didn't detect any difference between the two states. The other states are ordered similar to level II.

When compared to the level II graph, major paths and LASSO selected variables for m1-m6 are quite similar, indicating that on those modules, most BSP, ASP and ASF events in level II belong to BSPN, ASPN and ASFN in level III. The two noteworthy features are: 1. while m1-ASP was significantly correlated with final score in level II, m1-ASPN and m1-ASPB are not selected by LASSO as necessary variables in level III; 2. m3-ASFB is a significant negatively correlated variable, similar to m3-ASF in level II.

On the other hand, the level III model tells a very different story on m7-m10:

1. Most interaction states on m7 are no longer selected by LASSO for explaining the variance in the final course grade. Compared to level II, in which 4 states are selected with 3 being significant, only m7-NS is selected in level III, and the correlation is not significant. Meanwhile, m7-ASFN still serves as a “hub” connecting multiple major paths in all three cohorts.

2. BSPB and BSPN states on m8-m9 have different compositions between the three cohorts. On m8-m9, most BSP events in the top cohort belongs to BSPN, while for the bottom cohort a significant fraction of BSP events belong to BSPB. This seems to be a likely reason why m8-ASP in level III has a much weaker positive correlation compared to what was observed for m8-ASP in level II, since the current reference state, BSPN, is occupied by more students with higher course score.

3. Interaction states on m8-m10 differ significantly between top and bottom cohorts. With the current arrangement of states, the
bottom third cohort aggregated onto a “corridor” consisting of major paths between LS, NS, BSPB, and ASFB states extending from m8 to m10, “anchored” by several significant LASSO selected variables. In contrast, this corridor is almost empty for the top cohort, and less populated for the middle third cohort. The top cohort is mostly concentrated on BSPN and ASPN states between m8-m10, which are only sparsely occupied by the bottom cohort.

V. Discussion

A. Including more contextual information led to better descriptions of student behavior

Our analysis demonstrates that by increasing the amount of contextual information associated with each pass-fail event, we can obtain more informative and accurate descriptions of students’ online learning behavior.

The baseline regression model 2 reveals little more than the fact that high performing students pass more modules. By including the module number information, the level I model shows that passing modules m6-m9 are better indicators of higher final course score. However, it is hard to understand why failing on m10 is positively correlated with final course score. Note that the LASSO algorithm selected 17 out of 20 variables in this model, indicating that it has only limited ability to identify characteristic behavioral differences between students with high and low total course score.

The level II states added the contextual information on whether each pass-fail event happened before or after accessing the instructional materials. The level II model reveals that on modules m5, m7, m8 and m9, students with lower final score not only have lower passing rates, but are also more reluctant to access the instructional materials after repeated failure (LS and NS states). This could imply that those students either have less motivation to study or have otherwise lost confidence in their ability to learn from the IC. On the other hand, two observations are difficult to make sense of. First, the ASP states on m8, m9 and m10 are positively associated with final score, which implies that students with higher scores are more likely to fail their initial attempts and needed to study the IC.

Second, Figure 4 shows that many students in the bottom third cohort transitioned from NS and ASF states on m9 to BSP state on m10, which contains a harder problem than m9 in the AC.

The level III model included information on whether the pass-fail event was completed over a brief interval (less than 35 seconds). The addition of this information seems to be important for identifying characteristic behavioral differences between students with high and low final course scores, as it allows the LASSO algorithm to select only 20 out of 90 variables. The resulting model accounted for more variance in the final course score using 4 fewer variables than the level II model.

The level III parallel coordinate graph (Figure 5) shows a clear “corridor” from m8 to m10 for the bottom third cohort, consisting of major paths connecting either brief passing attempts (BSPB) or consecutive failed attempts without study (LS or NS). In contrast, the top third cohort mainly concentrated on normal passing attempts either before or after studying the IC (BSPN and ASPN) on the same modules, whereas the middle third cohort has more failed normal attempts (ASFN). Remarkably, for all three cohorts, the major paths between m8-m10 all originated from the same ASFN state on m7. This observation suggests that failing on m7 is not a characteristic difference between high and low performing students, but their different choices after experiencing the setback on m7 is: while the top and most of the middle cohort continued with learning (with the middle cohort being less successful), most of the bottom cohort gave up and resorted to guessing on the following modules.

The level III model also provides an explanation for the anomalous observations on level I and II models: many P and BSP events from the bottom 1/3 cohort on m9 and m10 are BSPB events (attempts completed in less than 35 seconds), while only a few students in this cohort studied the IC of the module.

B. Implications for Instructors

One of the important goals of learning analytics is to provide instructors with actionable recommendations to improve student learning. In that regard, the level III model is far superior to the other models.
The simple baseline model and level I model both rely on pass-fail events alone, which is similar to what is provided by many commercial online homework platforms. According to these two models, the average instructor can do little more than ask students to “work harder and pass more modules, especially on m6-m9”. The Level II model suggests that some students might have lost confidence toward the end, but the patterns are inconsistent. In addition, Level I and II models could mislead the instructor into believing that the bottom third cohort eventually mastered the content or even outperformed the top and middle cohorts on m9 and m10.

On the other hand, the level III model tells a more complete and more accurate story with three main takeaways:

1. On modules m1-m6, there are no qualitative differences in learning strategy for students with varying levels of ability to succeed in the course. In other words, almost everyone is trying to learn in the beginning.

2. Module 7 is challenging for most students as the instructional materials are insufficient for helping them learning how to solve the problems in the AC.

3. After experiencing a setback on m7, students with low course final scores are much more likely to employ a guessing strategy on the rest of the modules.

Given those takeaways, rather than telling students to “study harder” or “do better,” a more helpful message could be “Everybody experiences setbacks – it is alright to fail! The key to success is to not give up.” In addition, two interventions could potentially be beneficial for boosting students’ confidence:

1. Improve the quality of instruction on m7 to increase the chance of success especially for low performing students.

2. Conduct activities that develop a growth mindset, which has been shown to be beneficial for student success [42–44].

C. Implications for researchers conducting data-driven online learning research

First of all, we demonstrated that instead of employing more sophisticated algorithms, fine tuning different parameters, or using larger data sets, including detailed contextual information for each event analyzed can in some cases also be an effective approach for improving the accuracy and interpretability of learning analytics results.

Second, this study highlights the importance of the instructional design and platform capability in learning analytics. The contextual data that are crucial for the construction of the level II and III models are grounded in the unique OLM design blending assessment with instructional resources, which is made possible by the flexibility of the Obojobo platform. It is often the case that platform capability and instructional design can determine both the variety and accuracy of information that can be extracted from student log data [22,45], and in turn limits the depth of learning analytics. For example, the RISE project [46] is limited to simple analysis and visualization with limited contextual information, using data from generic online learning platforms. Therefore, it can be beneficial for all parties involved if data scientists and online learning researchers play a more actively role in the design, development or adoption of online learning platforms and online courses, rather than passively stay on the receiving end of educational data.

D. Caveats and future directions

One limitation of the current analysis is the use of a universal 35 seconds cutoff between Brief and Normal attempts. While this stringent criterion is favorable for avoiding false positives, it may not capture a significant number students who are not trying very hard on complex calculation problems that cannot be correctly solved within several minutes even by experts. This might explain why we still observe some students in the bottom cohort shift from Late Study and Abort states on m9 to the Brief Study Pass-Normal state on m10. In fact, for m9 and m10, exploratory data analysis [30] identified a separate distribution spending longer than average time solving the problem, while achieving a better correct response rate. Spending more than average time on those problems could be a characteristic behavior of the top 1/3 cohort just as "Brief" problem solving is characteristic for the bottom 1/3 cohort.

Another imperfection of the current analysis is that the scores on the OLM sequence is included in the total final course grade, which violates the conditions for linear regression. However, we think that this is a negligible small effect because:
1. the OLM sequence only accounts for 9% of the total grade and 2. all students received at least 90% of the score if they passed the module in 5 attempts. As a result, the failed states used in the linear model does not directly correlate to the module scores.

On a related issue, the total course final score contains exams, homework, lab and course participation scores, which makes it a (complex) measure of both content knowledge and effort. Future studies could explore the relation between students’ online learning behavior and their problem solving ability, conceptual knowledge or attitudes individually, using research validated instruments.

More importantly, in the current study, including more contextual information always led to better student model. But will it continue to be beneficial to include even more contextual information in the analysis? For example, when do students start working on the OLMs? Do they distribute the 10 modules over time, or do they try to complete all of them at once? How does the duration of student study sessions impact their ability to pass the assessment? How much time do students spend on solving each practice problem in the IC before accessing the solution? While including all those information could potentially lead to a more accurate description of students’ online learning behavior, will the benefit justify the cost for extracting those information from the log data? In addition, as the complexity of the model grows, it will also place a higher requirement on the sample size. Therefore, a valuable research question is whether there exists an optimum amount of contextual information that one could consider given a certain sample size.

Finally, the ultimate goal of learning analytics should not be limited to observing or predicting students’ learning behavior and learning outcome, but rather to inform the development of new interventions and evaluate their effectiveness. Therefore, a valuable future research direction is to examine whether the interventions inspired by the level III model could indeed lead to detectable change in the bottom cohort’s learning behavior, especially towards the end of the sequence.

VI. Acknowledgement
The authors would like to thank the Learning Systems and Technology team at UCF for developing the Obojobo platform. This research is partly supported by NSF Award No. DUE-1845436 and the Advancement of Early Career Researchers (AECR) Program at the University of Central Florida.

VII. References

[1] Kortemeyer G 2016 Work Habits of Students in Traditional and Online Sections of an Introductory Physics Course: A Case Study J. Sci. Educ. Technol. 25 697–703

[2] Kortemeyer G 2007 Correlations between student discussion behavior, attitudes, and learning Phys. Rev. Spec. Top. - Phys. Educ. Res. 3 1–8

[3] Formanek M, Buxner S, Impey C and Wenger M 2019 Relationship between learners’ motivation and course engagement in an astronomy massive open online course Phys. Rev. Phys. Educ. Res. 15 20140

[4] Lin S Y, Aiken J M, Seaton D T, Douglas S S, Greco E F, Thoms B D and Schatz M F 2017 Exploring physics students’ engagement with online instructional videos in an introductory mechanics course Phys. Rev. Phys. Educ. Res. 13 1–18

[5] Qiu J, Tang J, Liu T X, Gong J, Zhang C, Zhang Q and Xue Y 2016 Modeling and Predicting Learning Behavior in MOOCs Proceedings of the Ninth ACM International Conference on Web Search and Data Mining - WSDM ‘16 (New York, New York, USA: ACM Press) pp 93–102

[6] Kizilcec R F, Piech C and Schneider E 2013 Deconstructing Disengagement : Analyzing Learner
Subpopulations in Massive Open Online Courses Lak ’13 10

[7] Borrella I, Caballero-Caballero S and Ponce-Cueto E 2019 Predict and Intervene: Addressing the Dropout Problem in a MOOC-based Program Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale - L@S ’19 (New York, New York, USA: ACM Press) pp 1–9

[8] Arnold K E, Pistilli M D and Arnold K E 2012 Course Signals at Purdue: Using Learning Analytics to Increase Student Success 2nd Int. Conf. Learn. Anal. Knowl. 2–5

[9] You J W 2016 Identifying significant indicators using LMS data to predict course achievement in online learning Internet High. Educ. 29 23–30

[10] Li Y, Fu C and Zhang Y 2017 When and who at risk? Call back at these critical points Proceedings of the 10th International Conference on Educational Data Mining pp 168–73

[11] Jayaprakash S M, Moody E W, Lauria E J M, Regan J R and Baron J D 2014 Early Alert of Academically At-Risk Students: An Open Source Analytics Initiative J. Learn. Anal. 1 6–47

[12] Baker R S, Lindrum D, Lindrum M J and Perkowski D 2015 Analyzing Early At-Risk Factors in Higher Education e-Learning Courses Proc. 8th Int. Conf. Educ. Data Min. 150–5

[13] Gaševic D, Dawson S and Siemens G 2015 Let ’ s not forget : Learning Analytics are about Learning TechTrends71–64 59

[14] Winne P H and Hadwin A F 2017 Studying as self-regulated learning

[15] Wilcox B R and Pollock S J 2019 Investigating students ’ behavior and performance in online conceptual assessment Phys. Rev. Phys. Educ. Res. 15 20145

[16] Seaton D T, Kortemeyer G, Bergner Y, Rayyan S and Pritchard D E 2014 Analyzing the impact of course structure on electronic textbook use in blended introductory physics courses Am. J. Phys. 82 1186–97

[17] Alexandron G, Ruipérez-Valiente J A, Chen Z, Pedro J. Muñoz-Merino and Pritchard D E 2017 Copying @ Scale: Using Harvesting Accounts for Collecting Correct Answers in a MOOC Comput. Educ. 108

[18] Palazzo D J, Lee Y J, Warnakulasooriya R and Pritchard D E 2010 Patterns, correlates, and reduction of homework copying Phys. Rev. Spec. Top. - Phys. Educ. Res. 6

[19] Seaton D T, Bergner Y, Chuang I, Mitros P and Pritchard D E 2014 Who does what in a massive open online course? Commun. ACM 57 58–65

[20] Gutmann B, Gladding G E, Lundsgaard M and Stelzer T Mastery-style homework exercises in introductory physics courses: Implementation matters Phys. Rev. Phys. Educ. Res.

[21] Guthrie M W and Chen Z 2019 Comparing student behavior in mastery and conventional style online physics homework Physics Education Research Conference Proceedings 2019 ed Y Cao, S Wolf and M B Bennett (Provo, UT)

[22] Chen Z, Lee S and Garrido G 2018 Re-designing the Structure of Online Courses to Empower Educational Data Mining Proceedings of 11th International Educational Data Mining Conference ed K Elizabeth Boyer and M Yudelson (Buffalo, NY) pp 390–6

[23] Alexandron G, Ruipérez-Valiente J A, Chen Z, Muñoz-Merino P J and Pritchard D E 2017
Copying@Scale: Using Harvesting Accounts for Collecting Correct Answers in a MOOC Comput. Educ. 108

[24] Warnakulasooriya R, Palazzo D J and Pritchard D E 2007 Time to completion of web-based physics problems with tutoring. J. Exp. Anal. Behav. 88 103–13

[25] Barry C L, Horst S J, Finney S J, Brown A R and Kopp J P 2010 Do Examinees Have Similar Test-Taking Effort? A High-Stakes Question for Low-Stakes Testing Int. J. Test. 10 342–63

[26] Wise S L and Kong X 2005 Response Time Effort: A New Measure of Examinee Motivation in Computer-Based Tests Appl. Meas. Educ. 18 163–83

[27] Wise S L, Pastor D A and Kong X J 2009 Correlates of Rapid-Guessing Behavior in Low-Stakes Testing: Implications for Test Development and Measurement Practice Appl. Meas. Educ. 22 185–205

[28] Theobald E J, Aikens M, Eddy S and Jordt H 2019 Beyond linear regression: A reference for analyzing common data types in discipline based education research Phys. Rev. Phys. Educ. Res. 15 20110

[29] Tibshirani R 1996 Regression shrinkage and selection via the lasso J. R. Stat. Soc. Ser. B 58 567–288

[30] Garrido G, Guthrie M W and Chen Z 2019 How are students ’ online learning behavior related to their course outcomes in an introductory physics course? Physics Education Research Conference 2019 ed Y Cao, S Wolf and M B Bennett (Provo, UT)

[31] Center for Distributed Learning Obojobo

[32] Fakcharoenphol W, Potter E and Stelzer T 2011 What students learn when studying physics practice exam problems Phys. Rev. Spec. Top. - Phys. Educ. Res. 7 1–7

[33] Singh C and Rosengrant D 2003 Multiple-choice test of energy and momentum concepts Am. J. Phys. 71 607–17

[34] Chen Z Obojobo Sample Canvas Course

[35] Wickham H 2017 tidyverse: Easily Install and Load the “Tidyverse”

[36] R Core Team 2019 R: A Language and Environment for Statistical Computing Doc. Free. available internet http//www. r-project. org

[37] Schnipke D L and Scrams D J 1999 Modeling item response times with a two-state mixture model-a new approach to measuring speededness J. Educ. Meas. 34 213–32

[38] Chen Z, Garrido G, Berry Z, Turgeon I and Yonekura F 2018 Designing online learning modules to conduct pre- and post-testing at high frequency 2017 Physics Education Research Conference Proceedings (Cincinnati, OH: American Association of Physics Teachers) pp 84–7

[39] Prates M O and Cabral C R B 2009 mixsmsn: Fitting Finite Mixture of Scale Mixture of Skew-Normal Distributions Marcos J. Stat. Softw. 30 1–3

[40] Friedman J, Hastie T and Tibshirani R 2010 Regularization paths for generalized linear models via coordinate descent J. Stat. Softw. 33 1–22
[41] Zhao, Peng and B Y 2006 On model selection consistency of Lasso J. Mach. Learn. Res. 7 2541–63
[42] Yeager D S, Paunesku D, Walton G M and Dweck C S 2013 Excellence in Education: The Importance of Academic Mindsets Excell. Educ. Importance Acad. Mindsets 42
[43] Claro S, Paunesku D and Dweck C S 2016 Growth mindset tempers the effects of poverty on academic achievement Proc. Natl. Acad. Sci. U. S. A. 113 8664–8
[44] Dweck C S and Leggett E L 1988 A Social-Cognitive Approach to Motivation and Personality Psychol. Rev. 95 256–73
[45] Guthrie M W and Chen Z 2019 Adding duration-based quality labels to learning events for improved description of students’ online learning behavior. Proceedings of the 12th International Conference on Educational Data Mining ed M C Desmarais, C F Lynch, A Merceron and R Nkambou (Montreal, Canada: International Educational Data Mining Society (IEDMS))
[46] Bodily R, Nyland R and Wiley D 2017 The RISE framework: Using learning analytics to automatically identify open educational resources for continuous improvement Int. Rev. Res. Open Distance Learn. 18 103–22
[47] Dorea C C Y, Resende P A A and Gonçalves C R 2015 Comparing the markov order estimators AIC, BIC and EDC Transactions on Engineering Technologies: World Congress on Engineering and Computer Science 2014 (Springer Netherlands) pp 41–54
VIII. Appendix: Determining the Brief-Normal cutoff via skewed multi-mixture model.

Mixture model data fitting is conducted using the four different distribution models available in the R package mixsmsn: the normal distribution; the skew-normal distribution; the skew-Student-t distribution; and the skew-contaminated normal distribution (Skew-cn). The fitting algorithm searches for the optimum number of components and fitting parameters for each distribution model according to model selection criteria EDC, which is shown to be more reliable under certain conditions [47]. The four best fit models are then compared based on four model selection criteria: AIC, BIC, EDC and ICL. The model favored by the most criteria is adopted. If more than one model is favored, then the one favored by EDC is selected.

One challenge for data fitting is that problem solving durations can be as long as several thousand seconds, whereas Brief attempts are usually under 60 seconds. Therefore, the best-fit model may be selected because of a good fit for the long duration distribution but a less accurate fit for the shorter duration, or even not able to fit the short duration at all. To prevent this, we will only use the duration distribution below a maximum duration, and increase the maximum duration from 150s to 550s at 50s intervals to examine how the maximum duration affects the estimation of the Brief-Normal cutoff distribution.

Table A-1: Best fit model and the cutoff between the shortest and the next shortest distribution, for each maximum duration cutoff analyzed

| MAX DURATION | MODEL      | N COMP. | CUTOFF |
|--------------|------------|---------|--------|
| 150          | Skew.normal| 3       | 33.5   |
| 200          | Skew.cn    | 3       | 46.5   |
| 250          | Skew.cn    | 4       | 43.5   |
| 300          | Skew.cn    | 4       | 48.5   |
| 350          | Normal     | 4       | 30.5   |
| 400          | Normal     | 4       | 30.5   |
| 450          | Normal     | 5       | 30.5   |
| 500          | Normal     | 5       | 31.5   |
| 550          | Normal     | 5       | 30.5   |

The best fit model for each maximum duration, as well as the estimated cutoff between the first and second component, is listed in Table A-1. For maximum durations between 200s and 300s, the multi-component skew-cn distribution is selected to be the best fit model, with the 1st cutoff estimated at around 45 s. When maximum durations are more than 350s, the multi-component normal distribution is selected as the best fit model, with 1st cutoff at around 30s. However, the normal distributions run a higher risk of over fitting, since students' problem-solving duration distribution should be skewed by nature, as there is always a minimum amount of time required to solve any problem but no a clear upper limit. Therefore, we will take 35s as our "Brief-Normal" cutoff, which is close to the average of all the cutoffs obtained for different cutoffs. As shown in Figure A-1, the 35s cutoff sits right at the center of the first minimum of the distribution.

Figure A-1: Example of multi-component mixture model fit of the duration distribution, with maximum duration of 350s. The red line indicates the cutoff generated by the algorithm at 30s, and the blue line indicateds the average cutoff for all durations at 35s.