2023 Annual Conference & Exposition
Baltimore Convention Center, MD | June 25 - 28, 2023
The Harbor of Engineering Education for 130 Years
ASEE
Paper ID #37869

Board 204: A Trajectory-Clustering Framework for Assessing AI-Based Adaptive Interventions in Undergraduate STEM Learning

Dr. Mohammad Rashedul Hasan, University of Nebraska, Lincoln
Bilal Khan

A Trajectory-Clustering Framework for Assessing AI-based Adaptive Interventions in Undergraduate STEM Learning

Abstract

In this paper, we present a framework for quantifying the impact of interventions on the full trajectories of students' experiences. The interventions are given periodically based on student performance forecasting from an artificial intelligence (AI) model. We performed a small-scale randomized controlled trial for evaluating the impact of the AI-based intervention system on the undergraduate students of a science, technology, engineering, and mathematics (STEM) course. Intervention messaging content was based on machine learning forecasting models trained on data collected from the students in the same course over the preceding 3 years. Trial results show that the intervention produced a statistically significant increase in the proportion of students that achieved a passing grade. By applying the trajectory-analysis framework we find that the intervention impacts the stories of some types of students more than others, and use this to define new ways of identifying students who are most likely to benefit. Together these outcomes point to the potential and promise of just-in-time interventions for STEM learning and the need for larger fully-powered randomized controlled trials.

Introduction

Improving student retention in undergraduate science, technology, engineering, and mathematics (STEM) is critical for meeting the increased demand for STEM workforce in industry [1]. A key reason for the low retention rate [2] is students' poor academic performance, particularly in the first few years of college [3]. Recently, modern artificial intelligence (AI) methods such as machine learning (ML) have emerged as a low-cost approach for improving undergraduate STEM performance via sending incremental and contextually appropriate interventions [4, 5]. Specifically, this approach involves the creation of predictive ML models that use students' recent performance data (e.g., academic scores at the beginning of the semester) to forecast their summative performance, such as the final course outcomes. Such forecast messages are seen as interventions that serve to both inform students and motivate them to improve their academic performance [6, 4, 7]. The customizability and low implementation costs of AI-based solutions make them a potentially cost-effective, scalable approach for improving academic achievement, particularly in courses during the first two years of college, where STEM curriculum is fairly standardized, and performance is critical to long-term student retention [4, 7, 8].

And yet, despite the prospect of AI-based strategies for the delivery of interventions, there is a

scientific knowledge gap in our understanding of the mechanisms governing the efficacy of the interventions themselves [9, 10]. Indeed, little research has been done to investigate the variability in the effectiveness of just-in-time interventions among undergraduate STEM students. More specifically, the variability in the effectiveness of interventions on students at different performance levels is mostly unknown. For example, are all students who are not performing well equally likely to benefit? Which of them is *more* amenable to positive intervention impacts? Do these interventions only *positively* impact at-risk students, or could they impact students *negatively*, and if so, why? How do interventions impact not only the end outcomes but the intellectual journey that students experience? This paper presents a framework for answering these types of questions and applies it in the context of a small randomized efficacy trial.

Framework for evaluating the impact of intervention

The framework evaluates the impact of the intervention on the student's academic experience by focusing on (i) the student's summative performance and (ii) the student's academic experiential trajectory.

While the summative performance-based assessment enables only a reductive analysis (as it is based only on the student's final course grade), the experiential trajectory-based analysis allows quantifying the impact of an intervention on the full trajectories of the student's experience over the course of a semester. Thus, the latter approach can potentially help understand how the intervention impacts not just the end-of-the-semester summative performance, but also the evolving academic experiences or the "stories" of some types of students more than others. This understanding can be used to define new ways of identifying students who are most likely to benefit. Consequently, this could help develop fine-grained and effective just-in-time interventions.

To examine the variance in the impact of interventions on both the summative performance and the full academic trajectory, the framework utilizes a clinical trial [11], specifically a randomized controlled trial (RCT), which has been used as an effective clinical trial to study the impact of intervention [12, 13].

Influence of the intervention on the student's summative performance

This assessment involves measuring the impact of the intervention on the student's performance outcome. Specifically, it evaluates whether the interventions improved the student's course letter grade. Thus, this analysis concerns a single dimension of a student's academic experience, i.e., the letter grade.

Influence of the intervention on the student's experiential trajectory

On the other hand, the trajectory-based analysis concerns the student's full academic experience over the semester. It measures the impact of the intervention on each graded activity. Thus, this analysis is focused on a high-dimensional space of the student's academic experience.

For performing this analysis, the students are clustered after some time in the semester (e.g., a few weeks) has passed but before the first intervention is given. This pre-intervention clustering will

be done using the class assessment scores from the beginning of the semester until the time of clustering. The distinct clusters will enable identifying the students who share similar "initial" story types.

Similarly, a post-intervention clustering will be done at the end of the semester on a higher-dimensional space that includes all assessment scores from the beginning. The final clustering will identify groups of students with similar "final" story types based on their full academic experiential trajectories.

For evaluating the impact of the interventions, the pre-intervention and post-intervention clustering will be done for both the control and intervention groups. Then, the control and intervention cluster distributions will be compared to determine how the intervention influenced the stories of students for each initial story type. To quantify the impact over experiential trajectories, we will use the Shannon-Jensen divergence between the clusters of two groups.

Methods

The framework described in the previous section is used to conduct an RCT study in which each participant is randomly assigned to a group, and all the participants in the group receive (or do not receive) an intervention.

Study cohort

The study cohort consisted of 65 first-year students who enrolled in the introductory course on discrete structures at a large public university in the USA. The course had prerequisites of introductory programming and precalculus-level mathematics.

All students were informed at the beginning of class that they would be receiving 3 messages from "an AI" at a regular interval and that these 3 messages would contain "a forecast of your future performance". More specifically, the AI would determine if their prospects were "Good", "Fair", "Prone-to-Risk", "At-Risk", but in some cases the AI might declare that it was "Unable to make a prediction". Students were told that the messages they received would correspond to the final grades the AI had predicted for them, in accordance with Table 1.

Table 1: **Mapping of the AI's forecast to the AI's message.**

AI's forecast of final grade	**Message from AI**
$90\% \leq \text{grade} \leq 100\%$	"You are Good"
$80\% \leq \text{grade} < 90\%$	"You are OK"
$70\% \leq \text{grade} < 80\%$	"You are Prone-to-Risk"
$0\% \leq \text{grade} < 70\%$	"You are At-Risk"

The AI was instrumented by an ML app accessible through computer/cellphone browsers. Students were instructed to use their course management system (i.e., Canvas) credentials to log in to the app. When new forecasting was computed by the AI, the app notified the students by sending an automatic message to their email accounts. Students could check their forecasting messages by logging in to the app.

Randomized assignment

Just prior to week 6, the cohort was split into two groups. A randomly chosen one-half (33) of the students were assigned to the control group, and the remaining (32) were assigned to the intervention group. Students were not informed about the fact that a randomized assignment had taken place, or which group they were placed in.

All students (regardless of group) received 3 messages over the course of the semester, at the end of weeks $t = 6, 9$, and 12. The message at week 6 preceded the course midterm exam by 7 days; the message at week 12 preceded the final exam by 7 days; the message at week 9 was transmitted 1/3 of the way through the time interval between the midterm and final exams. These 3 timepoints were chosen in advance by contemplating natural pivots in the course delivery.

Intervention group

Within the intervention group, each student received a message at the end of week $t = 6, 9$, and 12. Each message indicated whether the AI had determined the student's prospects to be "Good", "Fair", "Prone-to-Risk", or "At-Risk". To generate the message for each student, data on the student's formal assessments (to date) were fed as input into an appropriate previously trained predictive ML model. The model's prediction was then translated into a message in accordance with Table 1. The AI-generated messages were delivered to each student via the app.

Control group

Within the control group, each student received a message at the end of week $t = 6, 9$, and 12. Each message stated that the AI had been "Unable to make a prediction" concerning the student's prospects in the course. These messages were delivered to each student via the app.

Predictive machine learning models

The predictive models described in this section were developed by the authors in previous work [5], and only a brief summary description is provided here. Three distinct predictive ML models were developed, $\mathcal{M}_6$, $\mathcal{M}_9$, and $\mathcal{M}_{12}$, to be used in determining intervention message content at weeks 6, 9, and 12, respectively.

Table 2: **Binning of the numerical final grade.**

Numerical final grade	**Label**
$90\% \leq$ grade $\leq 100\%$	"A"
$80\% \leq$ grade $< 90\%$	"B"
$70\% \leq$ grade $< 80\%$	"C"
$0\% \leq$ grade $< 70\%$	"Below C"

The **training data set** consisted of academic assessments collected from 537 students who were enrolled in the same course in the preceding six semesters. The number of cases in the training data set was thus 537. The dimensions of the data set were 17, consisting of 17 numerical predictors (numerical scores on homework assignments, quizzes, and exams), along with each

student's numerical final grade. The numerical grade was replaced with a categorical label of "A", "B", "C" or "Below C" using the binning scheme described in Table 2.

Table 3: **Outputs of the feature selection process.**

Model	Data Considered	Features selected
$\mathcal{M}_6$	Weeks 1 –6	Quiz 1 – 3 & Homework 1, 2
$\mathcal{M}_9$	Weeks 1 –9	Quiz 1 – 5 & Homework 1, 2, 3 & Midterm 1
$\mathcal{M}_{12}$	Weeks 1 –12	Quiz 1 – 7 & Homework 1, 2, 3, 4 & Midterm 1

Before building the model, **feature selection** was carried out by retaining only those predictors whose Pearson correlation with the final grade exceeded 0.45. The resulting features selected are presented in Table 3.

Rather than training a single classifier to predict all 4 class labels ("A", "B", "C", and "Below C") in one shot, we followed a hybrid approach [14]. Building each model $\mathcal{M}_t$ ($t = 6, 9, 12$) required the training of two classifiers:

$\mathcal{C}_t^1$ is a 3-label classifier that predicts either "A", "B" or "C and Below". It is trained on a transformed data set where the students who were labeled "C" or "Below C" have been relabelled as "C and Below". The distribution of the labels for this classifier is: A=252, B=156, and C and Below=129.

$\mathcal{C}_t^2$ is a binary classifier that predicts either "Below C" or "not Below C". It is trained on a transformed data set where the students who were labeled "A", "B", or "C" have been relabelled as "not Below C". The distribution of the labels for this classifier is: not Below C=396 and below C=141.

The predictions of the two classifiers $\mathcal{C}_t^1$ and $\mathcal{C}_t^2$ are combined to create the predictive model $\mathcal{M}_t$ as follows:

- If $\mathcal{C}_t^1$ predicts "A" or "B" then this output is taken to be the output of $\mathcal{M}_t$.
- If $\mathcal{C}_t^1$ predicts "C and Below" then $\mathcal{C}_t^2$ is consulted:
  - If $\mathcal{C}_t^2$ predicts "Below C" then that is taken to be the output of $\mathcal{M}_t$.
  - If $\mathcal{C}_t^2$ predicts "not Below C" then the output of $\mathcal{M}_t$ is "C".

This 2-stage design was chosen to address challenges associated with a lack of data and features, especially in early predictions (e.g., the $\mathcal{M}_6$ model). Performance measures for the $\mathcal{M}_6$, $\mathcal{M}_9$, and $\mathcal{M}_{12}$ models were described by the authors in their prior work [5], but are summarized in Table 9 within the Supporting Information section.

Measures of impact on student outcomes

We are interested in assessing the impact of the intervention on student performance outcomes. Towards this, we introduced the following measures:

P64. Percentage of students who passed (grade > 64)

P79. Percentage of students who earned a B or better (grade > 79)

P89. Percentage of students who earned an A or better (grade > 89)

Measures of Impact on Student Trajectories

While the outcome measures of the previous section allow us to capture whether the intervention made a summative difference "at the end", they fail to capture whether the intervention has an impact on the student's experiential trajectory over time. Thus, we are interested to quantify the extent to which the intervention impacted students' "stories" over the semester.

A typology of trajectories pre-intervention

Each student has had 5 assessments by the end of week 6, so we view each student's story up until that point as a time series of length 5, or as a point in $\mathbb{R}^5$. We considered the 65 student stories exhibited up to week 6 (i.e., the 65 points in $\mathbb{R}^5$), and performed clustering on these points.

Clusters were created based on the Euclidean distance in a 5-dimensional space. To identify the clusters, we used a Gaussian Mixture Model (GMM) technique [15]. GMM is a probabilistic model that assumes that the instances were generated from a mixture of several Gaussian distributions whose parameters are unknown. Each cluster generated by GMM can have a different ellipsoidal shape, size, density, and orientation. We select GMM because of its ability to identify arbitrary-shaped clusters. To determine the optimal number of clusters, we used a Bayesian GMM. This allowed us to infer an approximate posterior distribution over the parameters of a Gaussian mixture distribution. In the Bayesian model, we need to provide an upper bound for the number of clusters that we think the dataset might contain. Then, the model automatically assigns weights equal (or close) to zero to unnecessary clusters.

Through the above process, we found 4 clusters. We refer to these as the "initial" story types: I_1, I_2, I_3, and I_4. The 65 trajectories are illustrated in Fig. 1, with each being assigned one of 4 colors based on its initial cluster membership or initial story type.

A typology of trajectories post-intervention

By the end of the semester, each student had 17 assessments, and so their story in its entirety could be viewed as a time-series of length 17, or as a point in $\mathbb{R}^{17}$. We examined the 65 student stories exhibited over the whole semester (the 65 points in $\mathbb{R}^{17}$), and clustered these points in $\mathbb{R}^{17}$ following the procedure described above, i.e., using GMM and Bayesian GMM. The procedure led us to conclude that there were 3 "final" story types F_1, F_2, and F_3. The 65 trajectories are illustrated in Fig. 2, with each being assigned one of 3 colors based on its final cluster membership or final story type.

The histogram in Figure 3 shows the number of students having each of the 4 initial and 3 final story types within the control group. Figure 4 provides the corresponding data for the intervention group.

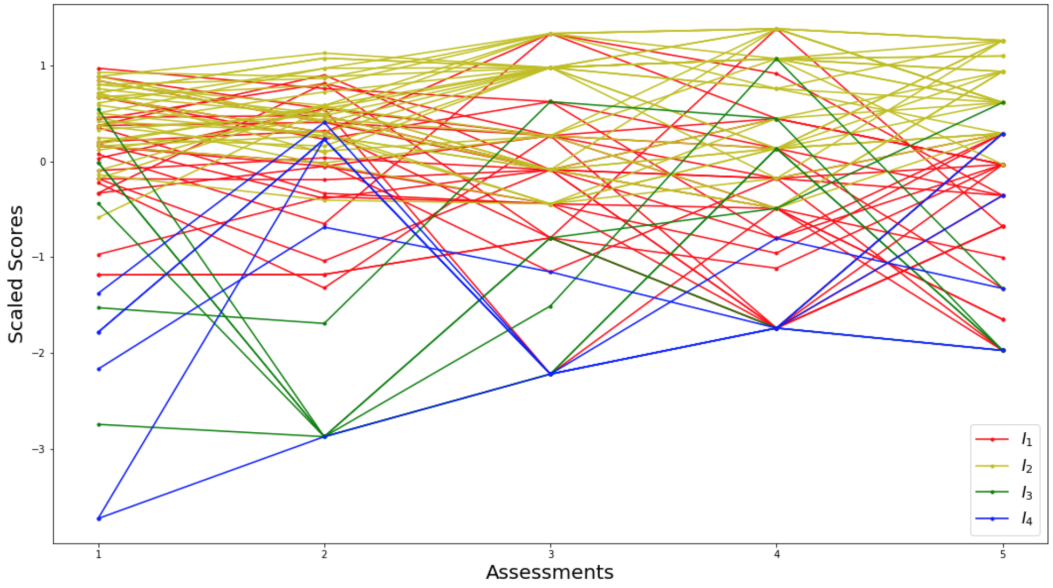

Figure 1: Entire cohort's trajectories at week 6 (color based on initial type)

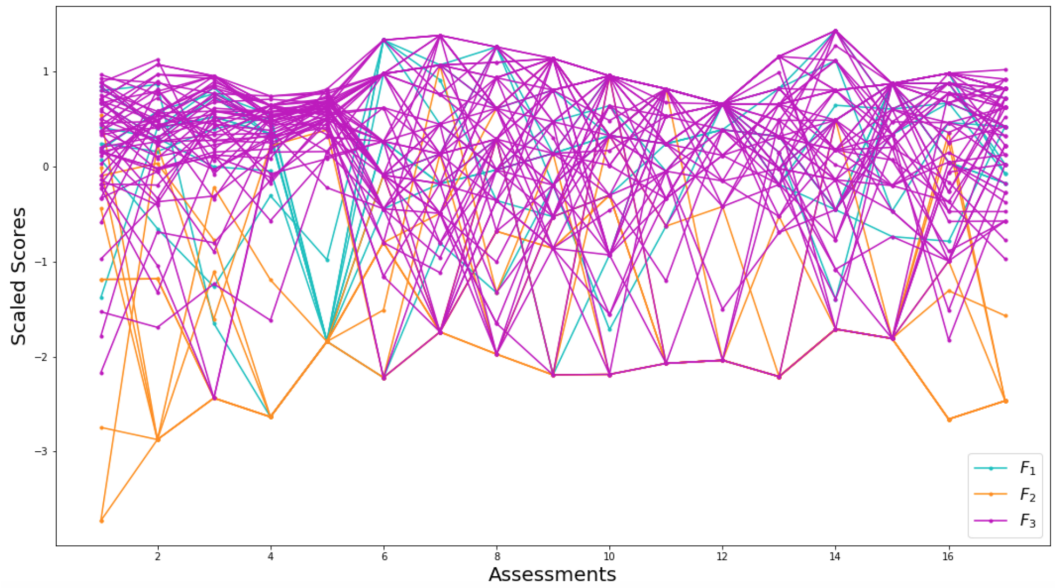

Figure 2: Entire cohort's trajectories at week 16 (color based on final type)

Measures of the impactfulness on trajectory types

Following the steps described previously, each of our 65 students was assigned an initial story type (I_1, I_2, I_3, or I_4) and a final story type (F_1, F_2, or F_3). For each initial story type I_k (for $k = 1, 2, 3, 4$), we computed the distribution over final story types (F_1, F_2, F_3) using data from students in the control group. Simultaneously, for that initial story type I_k, we computed the distribution over final story types using data from students in the intervention group. These two distributions were then compared by computing the Shannon-Jensen divergence between them.

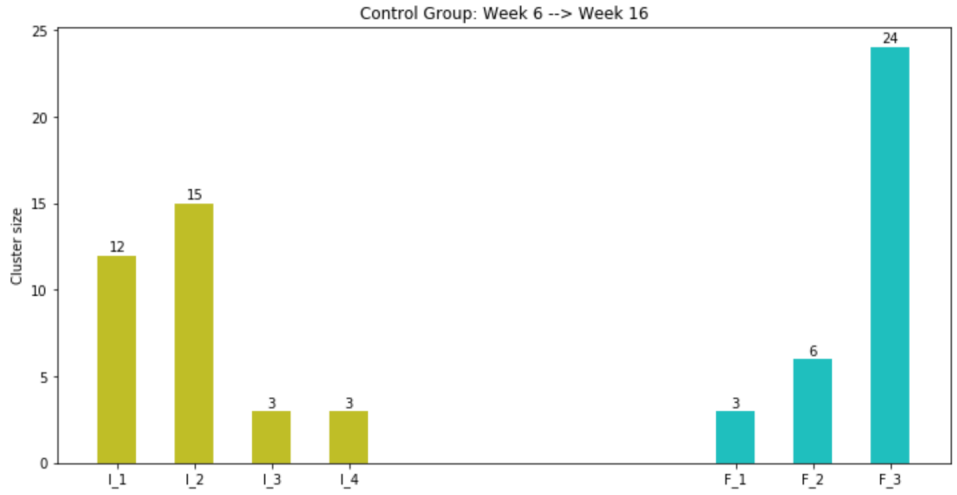

Figure 3: Clustering of the Control Group at weeks 6 and 16

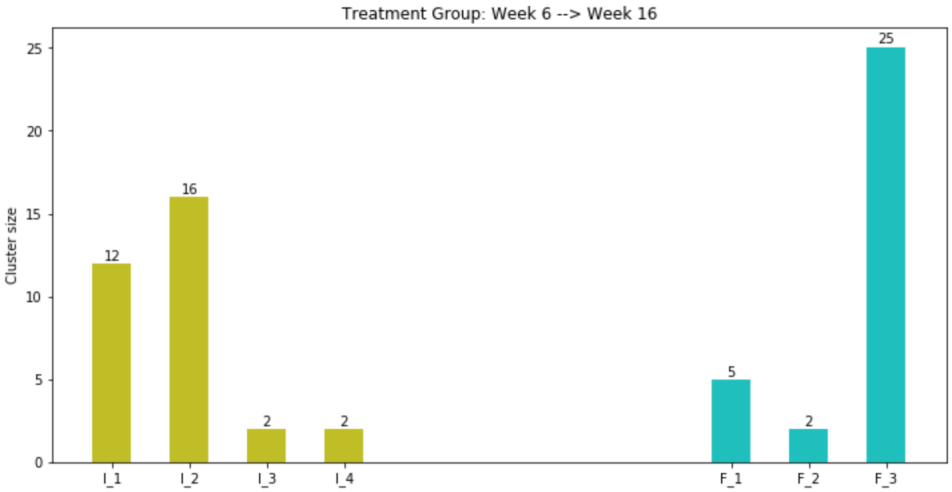

Figure 4: Clustering of the Intervention Group at weeks 6 and 16

This divergence value was taken to quantify the extent to which the intervention impacted the stories of students who started the course with an initial story type I_k. As it turns out, the intervention was not equally impactful in changing the type of story a student experiences; some types were more significantly impacted than others. The findings from this analysis are presented in the next section.

Results

Impact on student outcomes

Table 4 shows the number of students with final grades above and below thresholds of 64, 79, and 89 for the intervention and control groups.

Table 4: **Binomial Test Results for Three Threshold Outcomes.**

Threshold	Intervention		Control		p-value	Relative Risk
	# $\geq$ t	# $\leq$ t	# $\geq$ t	# $\leq$ t		
64	29	3	24	9	**0.013**	0.34
79	22	10	18	15	0.074	0.69
89	15	17	11	22	0.077	0.80

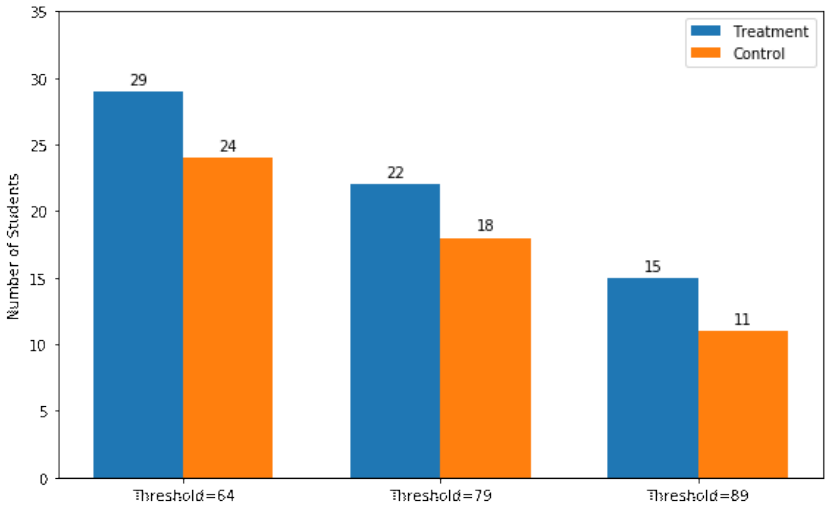

Figure 5: Outcome measures P64, P79, and P89 for Intervention & Control Groups

The number of students with a final grade above 64 (i.e., "passing" students) was higher in the intervention group (See Fig. 5), and this improvement was statistically significant at a 5% level (p-value = 0.013). For threshold 79 (the cutoff for a final grade of "B") and threshold 89 (the cutoff grade for a final grade of "A"), the p-values indicate a 7% probability of the observations being explained by pure chance; for these outcome measures, the impact of the intervention was detected only at the 10% significance level.

Additionally, Relative Risk (RR) was used to assess the substantive significance within the Randomized Controlled Trial (RCT). The RR for the thresholds 64, 79, and 89 were found to be 0.34, 0.69, and 0.80, respectively. Given that the RR values were all <1, the intervention was confirmed to reduce the number of students below the threshold relative to the control group.

Impact on student trajectories

All too often, the effectiveness of interventions is measured by a summative outcome variable that evaluates the impact on student performance (e.g., final grade). What is missing from such considerations is a sense of the trajectories by which students arrive at these summative endpoints. Capturing the distinct types of trajectories is a necessary first step to tailored intervention design.

Table 5 shows the transition probabilities from each type of initial story to each type of final story,

computed over the control group. For example, the table shows that in the control group, 8.3% of the students with initial story I_1 ended up having final story F_1.

Table 5: **Control Group's Transition Matrix (Week 6 $\rightarrow$ 16).**

		Final Story			
		Type F_1	**Type** F_2	**Type** F_3	**Entropy**
Initial Story	**Type** I_1	0.083	0.166	0.75	0.721
	Type I_2	0.066	0.0	0.933	0.244
	Type I_3	0.0	1.0	0.0	0.0
	Type I_4	0.333	0.333	0.333	1.098

Table 6 shows the corresponding transition probabilities for students in the intervention group. For example, the table shows that in the intervention group 16.7% of the students with initial story I_1 ended up having final story F_1—nearly twice the figure observed in the control group.

Table 6: **Intervention Group's Transition Matrix (Week 6 $\rightarrow$ 16).**

		Final Story			
		Type F_1	**Type** F_2	**Type** F_3	**Entropy**
Initial Story	**Type** I_1	0.167	0.0	0.833	0.450
	Type I_2	0.187	0.0	0.812	0.482
	Type I_3	0.0	0.5	0.5	0.693
	Type I_4	0.0	0.5	0.5	0.693

Table 7 presents the Shannon-Jensen divergence between the transition probability distributions exhibited by the intervention and control groups, for each of the initial story types. For example, the table shows that the intervention does not impact students with initial story type I_2 (SJ=0.130) as much as it does students with initial story types I_1, I_3, or I_4 (SJ=0.256, 0.464, 0.363, respectively). Given such a finding, one might seek to identify students who are likely to be impacted by the intervention (i.e., who are not of type I_2 by week 6). For simplicity, and as a proof of concept, we present a decision tree model trained to perform this binary classification task (type I_2 or type *not* I_2); the decision tree is shown in Figure 6. This decision tree can be used at Week 6 to identify students whose "stories are likely to be impacted". The tree is interpretable: the students most likely to be impacted are those whose score on Quiz 3 was below 6.5, and whose performance on Quiz 2 was below 3.5. Note that because our data set was quite small, we used the entire data to train the decision tree, in lieu of a training/testing split as is customary.

Table 7: **Shannon-Jensen Divergence of Intervention/Control Groups.**

		SJ Divergence
Initial Story	**Type** I_1	0.256
	Type I_2	0.130
	Type I_3	0.464
	Type I_4	0.363

The decision tree in Figure 6 seeks to identify students whose story type is likely to be disrupted by the intervention (as evidenced by high Shannon-Jensen divergence); it does not pass judgment on whether that disruption is beneficial to the student. To examine the nature of the disruption, in Table 8, we looked at the distribution of final grades for students of type I_2 versus *not* I_2, in both the control and treatment groups. We find that the mean final grade for students in group I_2 was negligibly different in the treatment versus control groups (when seen in the light of within-group variance). By comparison, the mean final grade for students in groups that were *not* I_2 was 16 points higher in the treatment group (70.40) compared to the control group (58.06). We wish to reiterate here that the trajectory clusters (both initial I_1, I_2, I_3, and final F_1, F_2, F_3, F_4) were all generated in an unsupervised manner based on similarities in longitudinal data, and the figures in Table 8 merely confirm that those trajectory groupings are consistent with improvements (for I_1, I_3, I_4) or stasis (for I_2) of the final grade. Unlike the previous outcome-based evaluation, the trajectory-based approach here has the added advantage of extracting distinct classes of student experiential trajectories ("stories"), each of which might be impacted to a varying degree by the intervention. Such a perspective could serve as a starting point for more nuanced future intervention designs.

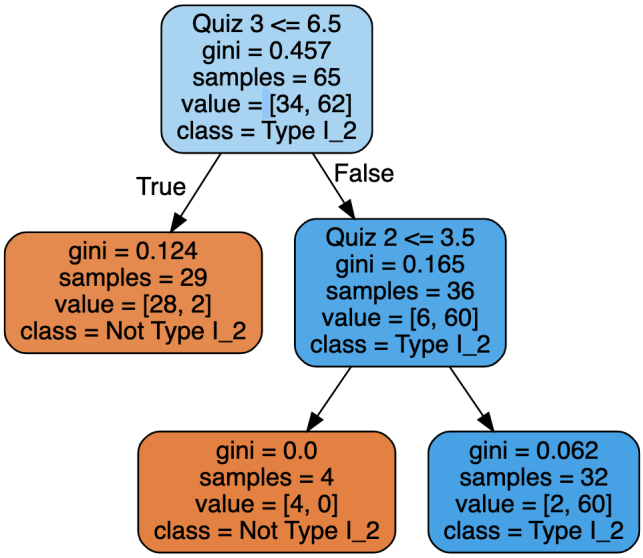

Figure 6: A decision tree for classifying students likely to be impacted by the intervention.

Table 8: **Average and standard deviation of the total score for both the control and treatment segments for the groups I_2 and not I_2.**

	Control			Treatment		
	Total Students	Avg.	Std.	Total Students	Avg.	Std.
I_2	15	91.03	9.63	16	90.06	10.27
Not I_2	18	58.06	26.93	16	70.40	24.24

Conclusions

We developed a framework to assess the impact of AI-based interventions on both the student's summative outcome and their full academic trajectory of experience. The framework is used for a small-scale RCT study. Results obtained from the study show the potential for AI-based interventions in improving STEM education.

- In terms of performance outcomes, we found that the intervention increased the number of students with a final grade above 64 (i.e., "passing" students) to a statistically significant extent. Although there was some improvement for students at other higher cutoffs, the study's sample size did not allow us to ascertain that such effects were not attributable to chance.
- We also introduced a new more holistic form of understanding outcomes. For this we introduced a method for clustering student trajectories, to determine distinct types of student stories. Building on this, we described a method to evaluate which types of stories are most significantly impacted by engaging in an intervention. Applying this general technique we found that our intervention was more impactful on students with an initial story *not* of type I_2, than those whose initial story was of type I_2.

Our research findings are limited by statistical power implications of the small cohort size (65 students), short duration (one semester), predictor granularity (17 timepoints), and intervention frequency (3 timepoints). A future RCT that is larger on any/all of these axes will allow us to test richer hypotheses, such as whether it is possible to cluster students based on their distal characteristics and proximal trajectories, towards the design of tailored interventions.

Supporting information

Performance of the predictive ML models. We used 80% of the N=300+ students' data to train the ML-based framework and tested it using 20% data. The performance of our predictive ML models is given in Table 9. The model used more features for the later predictions, as a result, the quality of the predictions improved. The model was tuned to increase precision and recall for the "You are At-Risk" group [5]. However, this improvement came at the cost of lower precision and recall for the "You are Prone-to-Risk" group.

Acknowledgments

This project was supported in part by a grant from the U.S. National Science Foundation (NSF DUE 2142558). The second author received support from the U.S. National Institutes of Health (NIH NIGMS P20GM130461 and NIH NIAAA R21AA029231), which facilitated the longitudinal analyses in this work.

Table 9: **Performance of the machine learning models.**

		Predictions by		
Class	**Measure**	$\mathcal{M}_6$ **at 6 wks**	$\mathcal{M}_9$ **at 9 wks**	$\mathcal{M}_{12}$ **at 12 wks**
At-Risk	Precision	0.70	0.79	0.92
	Recall	0.79	0.90	0.83
	F1	0.74	0.84	0.87
Prone-To-Risk	Precision	0.44	0.58	0.62
	Recall	0.38	0.52	0.76
	F1	0.41	0.55	0.68
Ok	Precision	0.68	0.74	0.81
	Recall	0.56	0.59	0.74
	F1	0.61	0.66	0.77
Good	Precision	0.66	0.76	0.84
	Recall	0.79	0.92	0.88
	F1	0.72	0.83	0.86
Overall Accuracy		0.64	0.73	0.80

References

[1] "Bureau of labor statistics," https://www.bls.gov/emp/tables/stem-employment.htm, accessed: 02-28-2021.

[2] A. Sithole, E. Chiyaka, P. McCarthy, D. Mupinga, B. Bucklein, and J. Kibirige, "Student attraction , persistence and retention in stem programs : Successes and continuing challenges," *Higher Education Studies*, vol. 7, no. 1, pp. 46–59, 2017. [Online]. Available: http://dx.doi.org/10.5539/hes.v7n1p46

[3] X. Chen, "Stem attrition among high-performing college students: Scope and potential causes," *Journal of Technology and Science Education*, vol. 5, no. 1, pp. 1–19, Jan. 2015.

[4] K. E. Arnold and M. D. Pistilli, "Course signals at purdue: Using learning analytics to increase student success," *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, pp. 267–270, 2012. [Online]. Available: http://doi.acm.org/10.1145/2330601.2330666

[5] M. R. Hasan and M. Aly, "Get more from less: A hybrid machine learning framework for improving early predictions in stem education," in *The 6th Annual Conf. on Computational Science and Computational Intelligence, CSCI 2019 (CSCI'19)*, 2019.

[6] D. F.-V. Nostrand and R. S. Pollenz, "Evaluating psychosocial mechanisms underlying stem persistence in undergraduates: Evidence of impact from a six-day pre-college engagement stem academy program." *CBE Life Sci Educ.*, vol. 16, no. 2, 2016.

[7] L. C. Page and H. Gehlbach, "How an artificially intelligent virtual assistant helps students navigate the road to college," *AERA Open*, vol. 3, no. 4, p. 2332858417749220, 2017. [Online]. Available: https://doi.org/10.1177/2332858417749220

[8] Y. Chen, A. Johri, and H. Rangwala, "Running out of stem: a comparative study across stem majors of college students at-risk of dropping out early," *LAK '18: Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, pp. 270–279, March 2018.

[9] G. Hansen, T. Brothen, and C. Wambach, "An evaluation of early alerts in a psi general psychology course," *The Learning Assistance Review*, vol. 7, no. 1, pp. 15–23, 2002.

[10] T. Brothen, C. Wambach, and N. Madyun, "Early alerts ii: An experimental evaluation," *Research and Teaching in Developmental Education*, vol. 20, no. 1, pp. 22–28, 2003.

[11] H. Nichols, "How do clinical trials work and who can participate?" Available at https://www.medicalnewstoday.com/articles/278779evidence, 2018, [Accessed on 04/12/2020].

[12] S. Freeman, S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt, and M. P. Wenderoth, "Active learning increases student performance in science, engineering, and mathematics," *Proceedings of the National Academy of Sciences*, vol. 111, no. 23, pp. 8410–8415, 2014. [Online]. Available: https://www.pnas.org/content/111/23/8410

[13] B. Styles and C. Torgerson, "Randomised controlled trials (rcts) in education research –methodological debates, questions, challenges," *Educational Research*, vol. 60, no. 3, pp. 255–264, 2018. [Online]. Available: https://doi.org/10.1080/00131881.2018.1500194

[14] M. Aly and M. R. Hasan, "Improving stem performance by leveraging machine learning models," in *the Proceedings of the International Conference International Conference of Frontiers in Education (FECS'19)*, 2019, pp. 205–2011. [Online]. Available: https://csce.ucmss.com/cr/books/2019/LFS/CSREA2019/FEC7082.pdf

[15] D. Reynolds, *Gaussian Mixture Models*. Boston, MA: Springer US, 2009, pp. 659–663. [Online]. Available: https://doi.org/10.1007/978-0-387-73003-5_196