7-21-2020

Variability in the Effectiveness of Psychological Interventions based on Machine Learning in STEM Education

Mohammad Hasan
hasan@unl.edu

Bilal Khan
University of Nebraska-Lincoln, bkhan2@unl.edu

Variability in the Effectiveness of Psychological Interventions based on Machine Learning in STEM Education

Mohammad Rashedul Hasan[1*], Bilal Khan[2],

1 Dept. of Computer Science and Engineering, University of Nebraska-Lincoln, USA
2 Dept. of Sociology, University of Nebraska-Lincoln, USA

* hasan@unl.edu

Abstract

This manuscript presents a framework to investigate the variability in the effectiveness of psychological interventions supported by Machine Learning (ML) based early-warning systems (EWS) in science, technology, engineering, and mathematics education. It emphasizes the importance of investigating the resulting variability and suggests that effective EWS cannot be designed without a deeper understanding of the variability. The framework uses an ML-based model to predict students' academic performance early in the semester for a Sophomore-level Computer Science course at a public university in the United States. The students were given psychological interventions by sending their end-of-term performance forecast thrice during the semester. A randomized control trial was designed to determine whether interventions made an overall positive impact on students' academic performance and whether there was variability in its impact. Results suggested that although interventions improved academic performance, they were not equally effective at different performance levels and that students at the same level reacted differently to these interventions.

1 Introduction

While the number of new jobs that require science, technology, engineering, and mathematics (STEM) knowledge is increasing in the United States of America (U.S.), the attrition rate in post-secondary STEM fields remains high [1–4]. A report published by the U.S. Department of Education's National Center for Education Statistics identified students' poor academic performance as the critical factor responsible for the high attrition rate [2]. The students' performance in the first few years of college was identified to be crucial for progression into subsequent years [5–9]. A large-scale systemic change was proposed to overcome the problem of poor academic achievement [1, 10]. However, such a solution would bring slow changes, cost a lot, and need to be tailored to individual institution's requirements [11]. Thus, there was an imminent need for a new cost-effective solution that required minimum systemic changes.

A feasible **solution** is to apply various types of interventions, such as active learning strategies to improve in class learning [12], light-touch interventions to improve learning outside the classroom [13], building STEM learning community to address both cognitive and social-psychological aspects of the learning process [14, 15]. The **psychological interventions** are an effective and inexpensive alternative that can be applied early during the semester [7, 11, 16]. It includes growth-mindset interventions

delivered via online sessions and early-warning interventions delivered by sending periodic warning messages. These interventions employ *nudges* to improve academic achievement, which relies on the analysis of human behavior, for example, habits, routines, and biases in normal decision-making [17]. Nudges can be used in an academic setting, for example, by sending an email to the student informing them of their end-of-term performance forecast [16] to improve academic achievement, and thus increasing the retention rate [11]. Social Cognitive Theory supports the **Early-warning systems** (EWS) and shows that students' non-cognitive psychological factors, such as motivation, play a critical role in improving their academic performance [18, 19].

The EWS requires student-profiles to deliver psychological interventions. Student test scores and cognitive factors have been used to create student-profiles as it correlates well with the student's performance [20–22]. The EWS that provide psychological interventions periodically throughout the semester needs to maintain dynamic student-profiles using cost-effective techniques. The recent advancements in Artificial Intelligence (AI) has made it possible to automate student profiling early during the semester [7, 23–27]. However, AI-based interventions require **Machine Learning** (ML) based predictive models. These models use students' current performance data, such as academic scores, at the beginning of the semester to predict what the student's performance (e.g., bad or good) will be at the end of the semester, thereby building student profiles automatically. The ML-based early intervention systems have emerged as a cost-effective and scalable solution to generate student profiles multiple times to increase students' motivation and engagement to improve their academic achievement and, thus, increasing the retention rate [7, 11, 28]. However, the ML-based approaches have not been used to study the variability of the resulted influence.

Most of the previous **ML-based predictive models** either predicted final numeric total scores, grades, or failure/pass status [25, 29]. For improving the predictive accuracy, these approaches used specialized grading systems such as standard-based grading. Uskov et al. [29] developed an ML-based mechanism that used students' academic performance as features to predict the final total scores or final grades. However, it only made one prediction based on all other features. Marbouti et al. [25] proposed an ML-based solution for making binary (at-risk or pass) periodic predictions. Students who obtained failing grades, i.e., lower grades than C, such as D, W, or F, were labeled as at-risk [27]. Its goal was to intervene and retain only at-risk students by preventing them from failing or dropping out of the course. However, this type of approach suffers from three limitations: (i) it is not enough to ensure that at-risk students obtain only passing grades for long-term retention [30], (ii) it is essential to make interventions to students who are forecasted to obtain B or C grade for increasing the graduation rate [31], and (iii) it uses standards-based grading, which is challenging to generalize across institutions. Thus, predictions at a fine-grained level were necessary to overcome these limitations.

Additionally, the **efficacy of the ML-based EWS** is not well understood [32, 33]. The variability in the effectiveness of interventions on students at different performance levels is mostly unknown. For example, it is not clear whether early interventions work only on students at the risk of failure or also on those who are not performing well but not necessarily at the risk of failing? Do these interventions only positively impact at-risk students and other groups, or could it impact students negatively and why? It is clear why some students become proactive after receiving an intervention while others do not?

There is no one-size-fit-for-all intervention to influence all students. Without any scientific understanding of these questions, the use of ML-based EWS as a generalized approach to improve undergraduate STEM education is likely to be unsuccessful. Thus, there was a **critical need** to investigate variability in the effectiveness of such early

interventions. The study **aims** to determine the variability in the effectiveness of psychological interventions given to the students early during the semester, and takes the first step towards building an effective EWS. We **hypothesized** that early interventions improve the academic performance of Computer Science undergraduate students (**Hypothesis 1**). We also hypothesized that variability exists in the effectiveness of the interventions (**Hypothesis 2**). Due to students' socioeconomic and psychological experiences it is possible that the predictions will have non-uniform influence at different performance levels as well as the same performance levels.

This novel research work **contributes** in two ways as follows:

- A framework is proposed to investigate the variability in the effectiveness of early interventions
- Knowledge is added in the area of automated early-warning systems using machine learning based predictive models

2 Methods

2.1 Intervention System

The ML-based framework by [23] was used to make periodic predictions at a fine-grained level for an undergraduate course. The model used current performance data of students to make predictions during the semester to assign them to one of the four groups in the future (by the end of the semester). Students were sent the predictions via a course management system. The instructor notified the students via an email when a prediction was released, as shown in Figure 1. Students were expected to log-in to the course management system to read their prediction, as illustrated in Figure 2. The automated ML-based prediction system to send interventions to students is described as follows:

2.1.1 Dataset

The final grading evaluation was based on weekly quizzes, homework assignments, midterm, and final exams. The performance data was collected from 472 students who were enrolled in the undergraduate Computer Science course between Fall 2015 and 2018. The predictions for the class of Fall 2019 were generated, which enrolled 65 students. The students were predicted to be in one of the four performance groups based on the criteria listed in Table 1:

Table 1. The labeling criteria for each class

Label	**Grade**	**Criteria**
Good	grade A	$\geq 90\%$
Ok	grade B	$80\% \leq \text{grade} < 90\%$
Prone-to-Risk	grade C	$70\% \leq \text{grade} < 80\%$
At-Risk	below grade C	$\text{grade} < 70\%$

2.1.2 Features

Course performance datasets are most effective when used as features for building predictive models [25]. Therefore, the performance data available to instructors for the

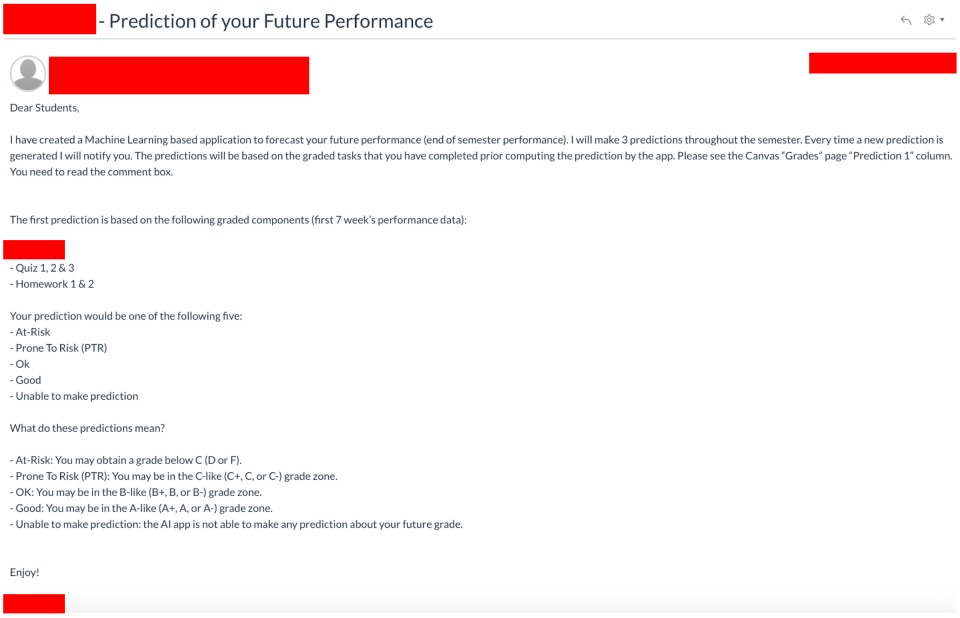

Fig 1. Email to the class notifying the release of a forecasted prediction

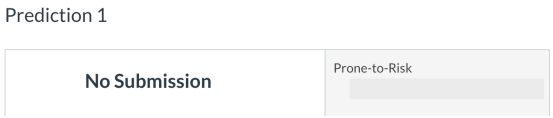

Fig 2. An example of a prediction forecast listed on the course management system

Table 2. Features and Prediction Timeline.

First prediction	Quiz 1 – 3 & Homework 1, 2	Week 1 – 6
Second prediction	Quiz 1 – 5 & Homework 1, 2, 3 & Midterm 1	Week 1 – 9
Third prediction	Quiz 1 – 7 & Homework 1, 2, 3, 4 & Midterm 1	Week 1 – 12

course under consideration were used as features. Exploratory Data Analysis was performed to select the features based on their correlation with the final grade was computed. Features with correlation values over 0.45 were used for training the models and generating three predictions, as described in Table 2.

2.1.3 Hybrid ML-based prediction

The ML-based framework in [23] addressed two challenges associated with a lack of data and features during early predictions. It made optimal classification when features were scarce. It did not perform four-class classification in a single step; instead, it singled out the groups successively in the order of their increasing importance. First, it identified the most critical group, i.e., at-risk group. Then, it identified the other three groups. The hybrid ML-based prediction framework is depicted in Figure 3.

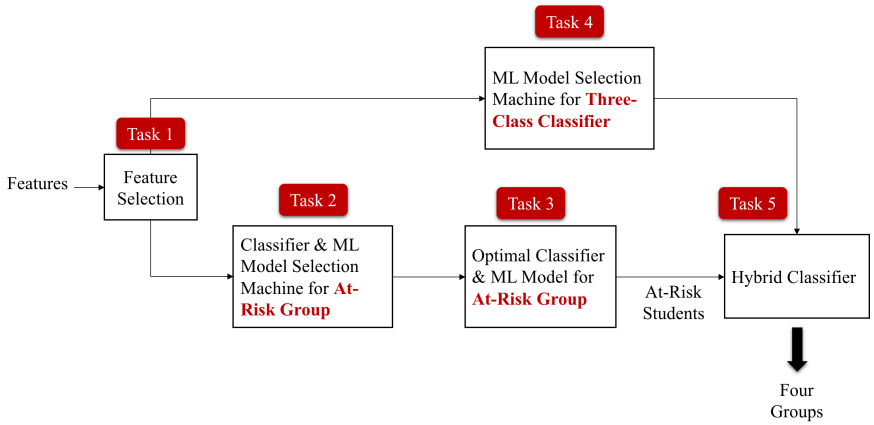

Fig 3. Hybrid ML Prediction Framework Pipeline [23]

The framework performed feature selection in Task 1, as described in 2.1.2. The selected features during Task 2 found the optimal four-class/binary classifier and the corresponding ML-based model. The classifier was selected based on the high recall and precision in the at-risk group. In Task 3, it predicted the at-risk students by using the optimal classifier in Task 2. The goal of Task 4 was to find the optimal ML model for three-class classification similar to [24] that predicted: class 1 (grade A), class 2 (grade B), and class 3 (grade C or below). In Task 5, the hybrid classifier took the three classes from Task 4 and the at-risk students from Task 3 and isolated the at-risk students from class 3 such that it only contained grade C students. The pipeline output was four predicted classes that ensured the optimality of the predictions of grade C group (prone to risk) and grade below C group (at-risk) based on Task 3 and 4. The pipeline in Figure 3 executed these tasks 1-5 for each prediction during the semester.

2.1.4 Number of Predictions

The ML-based hybrid framework made three predictions during the semester. The first prediction was made at week 6 for alerting students before the midterm exams. The second prediction was made at the end of week 9, to enable students to realize how their performance in the midterm and other tasks might influence their final grades. The third prediction was made at the end of week 12 for motivating students to prepare well for the final exam.

2.2 Randomized Control Trial (RCT)

To examine the variance in the impact of interventions, a clinical trial [34] is performed. While there exists various types of clinical trials, in education research the randomized controlled trials (RCTs) has been used as an effective clinical trial to study the impact of intervention [12, 35]. The RCT is an intervention study in which a group of subjects with similar characteristics are randomized to receive one of several defined interventions. It intends to find quantitatively the effect of an intervention on a defined outcome. It is a powerful tool for testing a hypothesis.

2.2.1 RCT Study Design

A RCT *parallel-group* study is designed in which each participant is randomly assigned to a group, and all the participants in the group receive (or do not receive) an

intervention. This study was conducted on a sophomore level STEM undergraduate major course in Computer Science at a large public university in US. Total 65 enrolled students took part in this study by signing an informed consent form. The study was approved by the university's Institutional Review Board (IRB #: 20180118001EX).

A flowchart of the methodology used in the study is shown in Figure 4. First, the ML-based predictive framework [23] generates performance predictions three times during the semester; on week 6, 9 and 12. These predictions are used to create early interventions. The intervention messages containing students' performance forecast are sent via the course management system (in Figure 2). On week 6, when the first prediction is generated, 50% students were randomly selected to provide interventions (32 students), while the remaining 50% (33 students) did not receive interventions. At the end of the semester the impact of the interventions was determined by performing a statistical significance test. In addition to this, the effect size of the interventions, confidence interval, statistical power as well as the variability in effectiveness of interventions were determined.

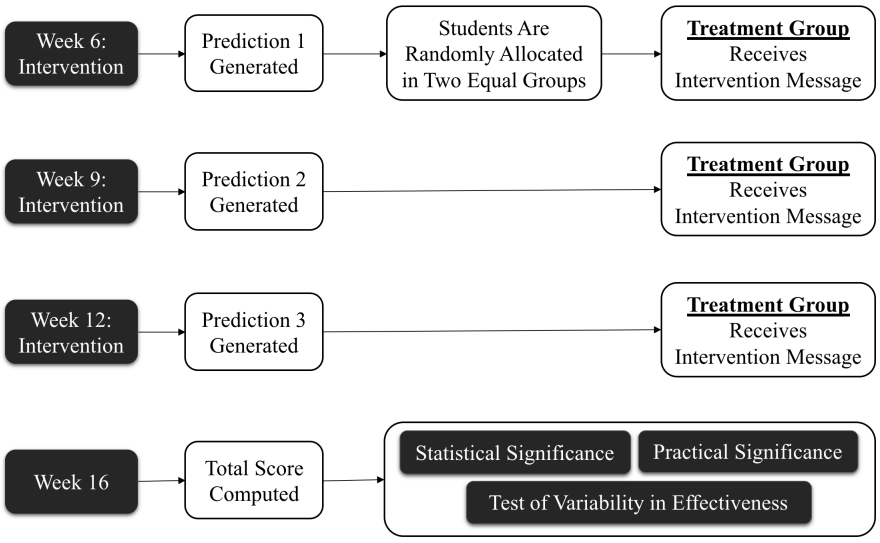

Fig 4. Randomized Control Trial (RCT) Parallel-Group Pilot Study Flowchart

2.2.2 RCT Study Evaluation

To evaluate the outcome of the RCT study both its statistical significance and practical significance are determined.

Test of Statistical Significance:
A statistical test was performed to determine whether the distribution of students into two categories (e.g., pass and fail) in the treatment group deviated significantly from the control group's distribution. In other words, the test was applied to assess whether the intervention increased the number of students above the threshold score significantly. The aim was to find whether the improvement in the treatment group (i.e., a higher number of students above the threshold score) was purely the result of a chance. The **one-tailed binomial test** was used as an exact test of the statistical significance of deviations due to the small sample size. The **null hypothesis** was formulated as follows:

- The difference in the distribution of students between the treatment and control groups is not statistically significant.

The null hypothesis is rejected if a significantly larger number of students is observed above the threshold in the treatment group as compared to that of the control group. The probability (p-value) was computed to obtain a total score greater than or equal to the threshold in the treatment group under the null hypothesis (i.e., based on the probability distribution of the control group). A 5% significance level is used (denoted by α) as the cut-off value to determine the probability of finding false negatives or making a Type I error (i.e., wrongly claim the there is an effect when there isn't). Thus is a p-value less than 5% is observed then there is less than 5% probability that any deviation from expected results (i.e., the distribution is according to the probability distribution of the control group) is due to chance only. In that case, we would reject the null hypothesis and conclude that the intervention made by the performance prediction app is statistically significant. We performed three one-tailed binomial tests for three threshold scores to determine the impact of the intervention at three performance levels.

Test of Practical Significance:
The practical significance of the results obtained from three one-tailed binomial tests was determined by using the following: **effect sizes, confidence interval and statistical power**.

As a measure of the point estimate of an effect size the risk ratio or relative risk is used. This metric is chosen because the study compared two groups (treatment and control) based on a dichotomous variable (e.g., pass vs. fail). Relative risk is computed by comparing the probabilities of group members being classified into one of the two categories (e.g., pass or fail) in both groups.

In addition, the precision of the effect sizes were determined by calculating respective confidence intervals. The confidence level is set at the standard value of 95%.

Finally, the statistical power of the study is computed, which provided the probability that the test correctly identified a genuine effect. In other words, it is the probability of rejecting a false null hypothesis or false negative (i.e., probability of not making the Type II error).

2.3 Investigating Improved Academic Performance

The impact of the intervention was examined using the weighted total score at the end of the semester. Three threshold score values closer to three critical cutoff grade points were chosen to see whether there was a significant increase in the number of students above the threshold scores. The threshold scores used were as follows:

- Passing grade cutoff: 64
- Letter grade B cutoff: 79
- Letter grade A cutoff: 89

A statistical significance test as well as practical significance test were performed to determine the impact of the intervention. The weighted total scores were computed at the end of the semester, using all graded components up to week 16. A score threshold was used to identify whether there is a significant increase in the number of students above the threshold score by performing a one-tailed binomial test. Three one-tailed binomial tests were performed for three threshold scores to determine the intervention's impact at three performance levels. Three binary distributions, i.e., pass or fail, $\geq$ grade "B" or $<$ "B", and $\geq$ grade "A" or $<$ "A" were explored.

2.4 Investigating Variability in the Impact of Predictions

The number of students belonging to four performance levels was counted, when the first prediction was sent at week 6 and at the end of semester in week 16. The weighted total score during these two times was used to determine student performance levels. The performance of the students from week 6 to 16 in treatment and control groups was assessed. The four performance-level clusters were determined using the score thresholds. These clusters can be *loosely* associated with the four performance groups used by the ML-based framework for generating predictions as previously mentioned: Cluster 1 $\rightarrow$ At-Risk, Cluster 2 $\rightarrow$ Prone-To-Risk, Cluster 3 $\rightarrow$ Ok, and Cluster 4 $\rightarrow$ Good.

- Cluster 1: Weighted Score < 64
- Cluster 2: Weighted Score ≥ 64 and < 79
- Cluster 3: Weighted Score ≥ 79 and < 89
- Cluster 4: Weighted Score ≥ 89

The following two tasks were performed to identify groups with variability in the effectiveness of the interventions. These tasks were conducted in both groups.

Task 1: To determine whether students at one performance level transitioned to other levels between Week 6 and 16. The transition probability matrix between the first and last prediction was computed. **Jensen-Shannon divergence** (JSD) metric [36] was used to quantitatively analyze the difference between the transition probabilities of four clusters between the two groups. It measured the similarity between transition probability distributions of the clusters for the treatment and the control groups.

Task 2: To determine the distribution of transitions from one performance level to other levels between Week 6 and 16. The uncertainty in the distribution in the transitions across the two predictions was computed. Entropy was used as a measure of uncertainty. Shannon's entropy [37] was used as it can measure the expected uncertainty of a random variable (COMMENT: citation needed. Also, I would add a block diagram showing the step-by-step procedure of the whole methodology followed in conduting your research.) (COMMENT HASAN: citation is provided and the flow-chart diagram is added in Figure 4)

3 Results

The hypothesis and variability in the effectiveness of interventions were evaluated. The accuracy of the predictions made by the proposed performance-prediction model was examined. It was a crucial step because the performance of the model could influence hypothesis validation. Low prediction accuracy may undermine the efficacy of the interventions.

3.1 Performance of the ML Model

The ML-based model generated three predictions during Week 6, 9, and 12. It was not directly possible to evaluate the predictions until the end-of-semester grades were obtained. Thus, 20% of the training data was used to evaluate the performance of the model.

In general, the model did not make highly accurate predictions at the beginning, which influenced the validation outcome of the hypothesis, as shown in Table 3. The model was tuned to increase precision and recall for the at-risk group [23]. However, it

Table 3. Performance of Three Predictions

		Prediction 1	**Prediction 2**	**Prediction 3**
At-Risk	Precision	0.70	0.79	0.88
	Recall	0.79	0.90	0.79
	F1	0.74	0.84	0.84
Prone-To-Risk	Precision	0.44	0.58	0.58
	Recall	0.38	0.52	0.71
	F1	0.41	0.55	0.64
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	0.78

Table 4. Statistical Significance Test Results at Three Threshold Scores

Threshold t	**Treatment:** $\#Students \geq t$	**Treatment:** $\#Students < t$	**Control:** $\#Students \geq t$	**Control:** $\#Students < t$	**p-value**
64	29	3	24	9	**0.013**
79	22	10	18	15	0.074
89	15	17	11	22	0.077

came at the cost of lower precision and recall for the prone-to-risk group. The model used more features for the later predictions, so the quality of the predictions improved.

3.2 Validation of Hypothesis 1: Interventions improve academic performance

The three threshold score values were chosen closer to three critical cutoff points: 64, 79, and 89. Table 4 shows the results obtained from the statistical significance test.

The number of students above the threshold 64 (i.e., "pass" students) was higher in the treatment group (Figure 5), and this improvement was statistically significant at a 5% significance level (p-value = 0.013).

For the threshold 79 (cutoff grade to determine students in the "B" grade region who were labeled as "Ok" students), the number of students above the threshold was higher in the treatment group (Figure 5). However, the p-value was 0.074, indicating that there was about 7% probability that the increase in the number of "Ok" students was purely a result of chance. This improvement was statistically significant at a 10% significance level. Similar observation was made for the threshold 89 that represented the cutoff grade for students in the "A" grade region who were labeled as "Good" students. The improvement was statistically significant at a 10% significance level (p-value = 0.077).

Additionally, the **substantive or practical significance** for the RCT was evaluated using the Relative Risk (RR), as a measure of the effect size. The RR for the three thresholds, i.e., 64, 79, and 89 scores, were 0.34, 0.69, and 0.80, respectively, as shown in Table 5. All three RR values were <1, indicating that the intervention reduced the number of students below the threshold compared to the control group. We

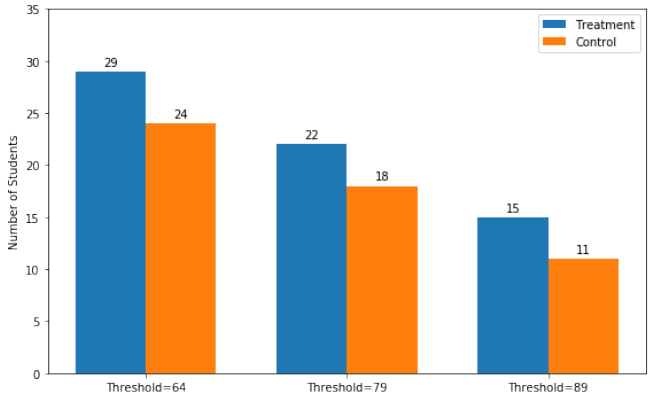

Fig 5. Number of Students above Three Thresholds (64, 79 & 89): Treatment & Control Groups

Table 5. Practical Significance Test Results at Three Threshold Scores

Threshold t	**Relative Risk**	**Confidence Interval**	**Power**
64	0.34	1.05	0.50
79	0.69	0.93	0.18
89	0.80	0.66	0.11

observe that the effect of intervention is the largest at the threshold 64 (lowest RR). In other words, the interventions are more effective to the reduce the number of failing students. However, the confidence interval (CI) for this effect is the highest. We observe that as the effect decreases for the other two thresholds, their CI reduces.

We also compute the power for three thresholds. We observe that intervention at threshold 64 has the highest power 0.50. However, at the other two thresholds the interventions are under-powered.

Thus, it was concluded that the improvement observed in the treatment group was not due to chance alone. **This conclusion validated our hypothesis 1**.

3.3 Validation of Hypothesis 2: Variability exists in the effectiveness of the interventions

The following two tasks were performed on both the treatment and control groups to validate this hypothesis. There are two related questions that we investigated.

- Does variability exist at different performance levels?
- Does variability exist at the same performance level?

Task 1: Determine whether students at one performance level transition to other performance levels between Week 6 and 16: Table 6 shows the cluster transition probability matrix between week 6 (rows) and week 16 (columns). In general, variability exited in the effectiveness of the interventions. This variability was more prominent among the Cluster 1, 2, and 3 (which were loosely associated with At-Risk, Prone-To-Risk, and Ok groups):

There was a 40% probability of students in Cluster 1 to improve their scores by moving to Cluster 2. However, these students showed a 60% probability to remain in the same cluster. There was a 25% probability of students in Cluster 2 to improve their scores by moving to Cluster 3. However, these students represented the highest tendency with a 75% probability to remain in the same cluster. There was a 36% probability of students in Cluster 3 to improve their scores by moving to Cluster 4. However, there was a 54% probability that these students remained in the same cluster.

Table 6. Treatment Group: Transition Probability Matrix (Week 6 → Week 16)

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Entropy**
Cluster 1	0.60	0.40	0.0	0.0	0.67
Cluster 2	0.0	0.75	0.25	0.0	0.56
Cluster 3	0.0	0.09	0.54	0.36	0.91
Cluster 4	0.0	0.08	0.0	0.92	0.29

The transition probability matrix from the treatment group was compared with the control group, as reported in Table 7. The **main difference** was in the non-zero values below the diagonal of the two matrices indicating performance degradation, i.e., the increased likelihood of moving from high- to low-performance clusters. The sum of the probabilities below the diagonal in the control group matrix was 56%, which was significantly larger than 17% probability in the treatment group. In the control group, the primary source of downward dragging was Cluster 2. The performance of students of Cluster 2 declined, as evident from their transition to Cluster 1 with a 43% probability. Also, an increased probability of performance decline in Cluster 4 was noticed in which students transition to Cluster 2 with a 13% probability.

Table 7. Control Group: Transition Probability Matrix (Week 6 → Week 16)

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Entropy**
Cluster 1	0.67	0.33	0.0	0.0	0.64
Cluster 2	0.43	0.29	0.29	0.0	1.07
Cluster 3	0 .0	0.0	0.55	0.44	0.69
Cluster 4	0.0	0.13	0.0	0.88	0.38

The JSD values, shown below, measures the similarity between transition probability distributions of the clusters for the treatment and the control groups.

- Cluster 1: 0.05
- Cluster 2: 0.45
- Cluster 3: 0.18
- Cluster 4: 0.05

The JSD was the highest between Cluster 2 of the two groups, which indicated that intervention made the most difference among Cluster 2 students of the treatment group. In general, there existed a tendency among students belonging to a cluster to remain in the same cluster. It was observed that interventions could disrupt this tendency, especially in low-performance clusters. However, it was not equally effective across all

performance clusters. Thus, there existed variability of the effectiveness in the interventions across **different** performance levels.

Task 2: Determine how distributed were the transitions from one performance level to other levels between Week 6 and 16: The last column of Table 6 showed the entropy of each row of the transition probability matrix in the treatment group. There was significant uncertainty in most of the performance levels. The highest uncertainty exists in Clusters 1 and 3. The students belonging to these two clusters did not transition to a single cluster with more than 40% probability. Students of these clusters diverge in their destination, albeit they received the same signal about their predicted performance for the end of the semester. Thus, there existed a variability even at the **same** performance level, meaning that students at the same performance level react differently to the intervention.

The results obtained from task 1 and 2 **validated our hypothesis 2**.

4 Student Feedback

A general user survey was conducted on the preliminary version of the proposed app at the end of the Fall 2019 semester and asked students about its usefulness and engagement with the app. About 87% of the students who used the app in Fall 2019 reported that the interventions helped improve their performance.

5 Discussion

The **work in this manuscript** used previously-obtained students' scores to forecast performance at a fine-grained level and to overcome a few limitations mentioned in [23]. For example, a student's current performance data (e.g., scores of the graded tasks) were used to predict a group, such as A, B, C, or below C grade, each student will belong to in the future. The periodic prediction of fine-grained performance level is expected to help students to track their future performance. For example, through periodic predictions, a grade B, or C, or at-risk student would know how their current efforts will shape their future performance and help them strategize efforts accordingly. RCT was designed to examine the variance in the impact of interventions [34]. RCTs have been previously used for similar purpose [12, 35]. A group of subjects with similar characteristics was randomized to receive one of several defined interventions. The pilot RCT used an ML-based predictive model of [23] to make predictions thrice early interventions during the semester at four performance levels. Only 50% of the students were randomly selected to provide interventions. The messages containing students' performance forecasts were sent via the course management system. The impact and the variability in the effectiveness of interventions were assessed at the end of the semester.

Making multiple predictions at a fine-grained level was challenging. Specifically, it was difficult to make optimal predictions at an early stage of the semester when student performance scores were scarce. An ML-based classifier could make accurate classification if a large number of datasets were used to train the classifier with many informative features. However, in a typical university course, an instructor does not usually have much historical data for training a classifier. A course could be taught by multiple instructors, who may use different evaluation techniques or difficulty level. Thus, a normalized set of historical data was not available, and students' academic scores were limited at the beginning of the semester.

A high entropy was observed in the transition from Cluster 1 in the treatment group. Observing high entropy in Cluster 1 was contrary to our expectation because this Cluster contained students who were at a high risk of failure. It showed that the interventions did not work for 60% of students in Cluster 1 (Table 6).

Although the entropy of Cluster 2 was slightly lower than that of Cluster 1, the strong inertia of Cluster 2 students to remain in the same Cluster (with 75% probability) was surprising. Cluster 2 students obtained scores between $64 \geq$ and < 79. Most of these poorly-performing students did not react to the intervention positively, which could be due to inaccurate forecasting received by these students.

It was expected that the entropy of Cluster 4 would be low because this Cluster contained students who obtained scores > 89. However, it was not clear why there exists high entropy in Cluster 3 that contained students mostly in the "B" grade range (scores between $79 \geq$ and < 89).

We believe that a more accurate forecasting model might smooth out some inconsistencies. However, it might not account for the varying impact of the interventions. In other words, there might be some intrinsic factors (e.g., socio-psychological background of students) that may contribute to this variability. The interventions are given to the students without considering the possible intrinsic factors. We conjecture that by customizing the interventions based on the intrinsic factors, it may be possible to reduce the observed variability and thereby to increase the impact of interventions.

6 Conclusion and Future Work

This article emphasized the importance of investigating the variability in the effectiveness of interventions generated by the ML-based early-warning systems in STEM undergraduate education. The ML-based forecasting models could identify poorly performing students early during the semester. These identified students could be given early interventions by sending forecasts of their future performance to help improve scores. Due to the low implementation cost, these EWS are easily scalable nationwide to build a competitive STEM workforce. Despite the promise these systems offer, there is a lack of understanding of their efficacy.

A framework was built to investigate the variability in the effectiveness of early interventions. As part of this framework, a randomized control trial was designed. The results showed that while interventions make an overall positive impact on students' academic performance, there is variability in its impact. We found that interventions at different performance levels are not equally effective and that students at the same level react differently to the intervention.

6.1 Conclusions

Two conclusions were drawn from this research work as follows:

- Early interventions can improve academic performance.
- There exists variability in the effectiveness of interventions, i.e., students from the same performance level do not react to the same intervention message coherently and, therefore, not equally benefitted.

6.2 Future Work

Further investigations are necessary to understand the variability in the effectiveness of interventions. We plan to conduct a clinical trial with a larger sample size to increase

statistical power Besides, it would be useful to identify the hidden factors that cause the variability in the effectiveness of the interventions, such as noncognitive factors, and the impact of such interventions.

References

1. Sithole A, Chiyaka ET, McCarthy P, Mupinga DM, Bucklein BK, Kibirige J. Student Attraction , Persistence and Retention in STEM Programs : Successes and Continuing Challenges. Higher Education Studies. 2017;7(1):46–59.

2. Chen X, Soldner M. STEM attrition: college students' paths into and out of STEM fields. Statistical Analysis Report. National Center for Education Statistics, Institute of Education Sciences, US Department of Education. 2013; p. 1–104. doi:https://nces.ed.gov/pubs2014/2014001rev.pdf.

3. National Science Board. Revisiting the STEM workforce: A companion to science and engineering indicators. Arlington, VA: NSB Publication; 2015. Available from: `https://www.nsf.gov/nsb/publications/2015/nsb201510.pdf`.

4. Adamuti-Trache M, Andres L. Embarking on and Persisting in Scientific Fields of Study: Cultural capital, gender, and curriculum along the science pipeline. International Journal of Science Education. 2008;30(12):1557–1584. doi:10.1080/09500690701324208.

5. Chen Y, Johri A, Rangwala H. 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. 2018; p. 270–279. doi:10.1145/3170358.3170410.

6. Chen X. STEM attrition among high-performing college students: Scope and potential causes. Journal of Technology and Science Education. 2015;5(1):1–19. doi:http://dx.doi.org/10.3926/jotse.136.

7. Arnold KE, Pistilli MD. Course Signals at Purdue: Using Learning Analytics to Increase Student Success. Proceedings of the 2nd International Conference on Learning Analytics and Knowledge. 2012; p. 267–270. doi:10.1145/2330601.2330666.

8. Seymour E, Hewitt NM. Talking about leaving: Why undergraduates leave the sciences). Boulder, CO: Westview Press; 1997.

9. Hunter AB. Why Undergraduates Leave STEM Majors: Changes Over the Last Two Decades. In: Seymour E, Hunter AB, editors. Talking about Leaving Revisited : Persistence, Relocation, and Loss in Undergraduate STEM Education. Cham, Switzerland: Springer Nature Switzerland AG; 2019. p. 87–114.

10. Fry CL. Achieving Systemic Change: A Sourcebook for Advancing and Funding Undergraduate STEM Education, The Coalition for Reform of Undergraduate STEM Education. The Association of American Universities. 2013; p. 1–30.

11. Nostrand DFV, Pollenz RS. 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. 2016;16(2). doi:https://doi.org/10.1187/cbe.16-10-0294.

12. Freeman S, Eddy SL, McDonough M, Smith MK, Okoroafor N, Jordt H, et al. Active learning increases student performance in science, engineering, and mathematics. Proceedings of the National Academy of Sciences. 2014;111(23):8410–8415. doi:10.1073/pnas.1319030111.

13. Rodriguez F, Rivas MJ, Matsumura LH, Warschauer M, Sato BK. How do students study in STEM courses? Findings from a light-touch intervention and its relevance for underrepresented students. PLOS ONE. 2018;13(7):1–20. doi:10.1371/journal.pone.0200767.

14. Solanki S, McPartlan P, Xu D, Sato BK. Success with EASE: Who benefits from a STEM learning community? PLOS ONE. 2019;14(3):1–20. doi:10.1371/journal.pone.0213827.

15. Cohen GL, Garcia J, Apfel N, Master A. Reducing the Racial Achievement Gap: A Social-Psychological Intervention. Science. 2006;313(5791):1307–1310. doi:10.1126/science.1128317.

16. Paunesku D, Walton GM, Romero C, Smith EN, Yeager DS, Dweck CS. Mind-Set Interventions Are a Scalable Treatment for Academic Underachievement. Psychological Science. 2015;26(6):784–793. doi:10.1177/0956797615571017.

17. Thaler RH, Sunstein CR. Nudge: Improving Decisions about Health, Wealth, and Happiness. New Haven: Yale University Press; 2008.

18. Cromley JG, Perez T, Kaplan A. Undergraduate STEM Achievement and Retention: Cognitive, Motivational, and Institutional Factors and Solutions. Institute of Education Sciences (ED); National Science Foundation (NSF). 2016;3(1):4–11. doi:https://doi.org/10.1177/2372732215622648.

19. Bandura A. Social cognitive theory of mass communication. Media Psychology. 2001;3:265–299.

20. Zhang G, Anderson TJ, Ohland MW, Thorndyke BR. Identifying Factors Influencing Engineering Student Graduation: A Longitudinal and Cross-Institutional Study. Journal of Engineering Education. 2004;93(4):313–320. doi:10.1002/j.2168-9830.2004.tb00820.x.

21. Jones BD, Paretti MC, Hein SF, Knott TW. An Analysis of Motivation Constructs with First-Year Engineering Students: Relationships Among Expectancies, Values, Achievement, and Career Plans. Journal of Engineering Education. 2010;99(4):319–336. doi:10.1002/j.2168-9830.2010.tb01066.x.

22. Van Soom C, Donche V. Profiling First-Year Students in STEM Programs Based on Autonomous Motivation and Academic Self-Concept and Relationship with Academic Achievement. PLOS ONE. 2014;9(11):1–13. doi:10.1371/journal.pone.0112489.

23. Hasan MR, Aly M. 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.

24. Aly M, Hasan MR. Improving STEM Performance by Leveraging Machine Learning Models. In: the Proceedings of the International Conference International Conference of Frontiers in Education (FECS'19); 2019. p. 205–2011. Available from: https://csce.ucmss.com/cr/books/2019/LFS/CSREA2019/FEC7082.pdf.

25. Marbouti F, Diefes-Dux HA, Madhavan K. Models for Early Prediction of At-risk Students in a Course Using Standards-based Grading. Comput Educ. 2016;103(C):1–15. doi:10.1016/j.compedu.2016.09.005.

26. Essa A, Ayad H. Student Success System: Risk Analytics and Data Visualization Using Ensembles of Predictive Models. Proceedings of the 2nd International Conference on Learning Analytics and Knowledge. 2012; p. 158–161. doi:10.1145/2330601.2330641.

27. Macfadyen LP, Dawson S. Mining LMS data to develop an early warning system for educators: A proof of concept. Computers and Education. 2010;54(2):588 – 599. doi:https://doi.org/10.1016/j.compedu.2009.09.008.

28. Page LC, Gehlbach H. How an Artificially Intelligent Virtual Assistant Helps Students Navigate the Road to College. AERA Open. 2017;3(4):2332858417749220. doi:10.1177/2332858417749220.

29. Uskov VL, Bakken JP, Byerly A, Shah A. Machine Learning-based Predictive Analytics of Student Academic Performance in STEM Education. In: 2019 IEEE Global Engineering Education Conference (EDUCON); 2019. p. 1370–1376.

30. Kovacs K. When a C Isn't Good Enough; 2016. Available at `https://www.insidehighered.com/news/2016/09/23/students-who-earn-cs-gateway-courses-are-less-likely-graduate-new-data-sh`

31. Baldasare A, Vito M, Del Casino Jr VJ. When a B Isn't Good Enough; 2016. Available at `https://www.insidehighered.com/views/2016/11/15/developing-metrics-and-models-are-vital-student-learning-and-retention-es`

32. Hansen G, Brothen T, Wambach C. An evaluation of early alerts in a PSI general psychology course. The Learning Assistance Review. 2002;7(1):15–23.

33. Brothen T, Wambach C, Madyun N. Early alerts II: An experimental evaluation. Research and Teaching in Developmental Education. 2003;20(1):22–28.

34. Nichols H. How do clinical trials work and who can participate?; 2018. Available at `https://www.medicalnewstoday.com/articles/278779#evidence`.

35. Styles B, Torgerson C. Randomised controlled trials (RCTs) in education research –methodological debates, questions, challenges. Educational Research. 2018;60(3):255–264. doi:10.1080/00131881.2018.1500194.

36. Lin J. Divergence measures based on the Shannon entropy. IEEE Transactions on Information Theory. 1991;37(1):145–151. doi:10.1109/18.61115.

37. Shannon CE. A mathematical theory of communication. The Bell System Technical Journal. 1948;27(3):379–423.