Effective Evaluation of Online Learning Interventions with Surrogate Measures

Anonymous
Anonymous Institution
anonymous@anonymous.edu

ABSTRACT

There is a growing need to empirically evaluate the quality of online instructional interventions at scale. In response, some online learning platforms have begun to implement rapid A/B testing of instructional interventions. In these scenarios, students participate in series of randomized experiments that evaluate problem-level interventions in quick succession, which makes it difficult to discern the effect of any particular intervention on their learning. Therefore, distal measures of learning such as posttests may not provide a clear understanding of which interventions are effective, which can lead to slow adoption of new instructional methods. To help discern the effectiveness of instructional interventions, this work uses data from 26,060 clickstream sequences of students across 31 different online educational experiments exploring 51 different research questions and the students' posttest scores to create and analyze different proximal surrogate measures of learning that can be used at the problem level. Through feature engineering and deep learning approaches, next problem correctness was determined to be the best surrogate measure. As more data from online educational experiments are collected, model based surrogate measures can be improved, but for now, next problem correctness is an empirically effective proximal surrogate measure of learning for analyzing rapid problem-level experiments.

Keywords

Surrogate Measures, Measures of Learning, A/B Testing, Educational Experiments

1. INTRODUCTION

There is a growing need to empirically evaluate the quality of online instructional interventions at scale. This is in part motivated by the lack of empirical evidence for many existing interventions, especially in mathematics. According to Evidence for ESSA, a website that tracks empirical research on educational practices created by the Center for Research and Reform in Education at Johns Hopkins University School of Education, only four technology based interventions have strong evidence for improving students' mathematics skills [5]. In response, more and more online learning platforms are creating infrastructure to run randomized controlled experiments within their platforms [21, 12, 20] in order to increase the impact of the their programs on student learning and facilitate research in the field. This infrastructure allows for rapid A/B testing of different instructional interventions. In an A/B testing scenario, students assigned to particular assignments or problems within these online learning platforms will be automatically randomized to one of multiple experimental conditions in which different instructional interventions will be provided to them. While this paradigm allows for rapid testing of many hypotheses, this rapid testing environment makes statistical analysis difficult. In some cases, students participate in many randomized controlled experiments in parallel or in quick succession. For example, in ASSISTments, an online learning platform in which students complete pre-college level mathematics assignments [9], students can be randomized between different instructional interventions for each mathematics problem in their assignment. In these scenarios, it is important to evaluate the effect of the interventions as quickly as possible. If one were to wait until the end of a section of the curriculum, or even the end of the current assignment before evaluating students' mastery of the subject matter, then the effect of an intervention for a single problem near the beginning of the assignment would be obfuscated by the effects of all the following interventions. For this reason, prior work has only used students' behavior on the problem they attempted after receiving an intervention but before receiving another intervention to evaluate the effectiveness of the first intervention [13, 17]. However, the measures used in prior work were chosen based on theory, without any empirical evidence that they are in fact an effective surrogate measure of learning.

To address the lack of empirical evidence for these proximal surrogate measures of learning, the first goal of this work was to create a variety of surrogate measures from students' clickstream data on the problem they attempted after receiving an experimental intervention. These measures were created through feature engineering, discussed in Section 3, and model fitting, discussed in Sections 4.1 and 4.2.

After creating surrogate measures, The second goal of this work was to evaluate how effective these measures were at estimating the treatment effects between pairs of conditions in online experiments. To achieve this goal, data was collected to compare 51 different pairs of conditions from 31 assignment-level online experiments with posttests in which students were exposed to the same intervention multiple times within the same assignment, but were not exposed to any other interventions. By determining the extent to which each measure was a surrogate for students' posttest scores, discussed more in Sections 2.3 and 4.4, the surrogate measures could be compared to each other.

After determining which surrogate measure was most suited for use in rapid online experiments, the third goal of this work was to explore the effects of using the chosen surrogate to analyze the results of online education experiments compared to using posttest scores to analyze the results, discussed more in Section 4.5.

To summarise, this work strives to answer the following three research questions:

1. What surrogate measures can be created from short sequences of students' clickstream data?
2. Which of these surrogate measures is the best surrogate for posttest score?
3. How does using this surrogate measure to analyze online educational experiments effect their results?

2. BACKGROUND

2.1 Rapid Online Educational Experimentation

Experimentation is a cornerstone of formative improvement of online instructional interventions [20, 1]. When making decisions about implementing changes to online learning programs, designers must understand which features will have the greatest impact on student learning. A/B testing, i.e., comparing students' performance when they are randomly exposed to different variants of feature, allows researchers to estimate the causal effect of a specific feature. This causal estimate can be used to determine which variant of a feature should be scaled system wide and inform design decisions for future product development.

Systems like ASSISTments E-TRIALS were established to allow researchers to test learning theories and feature ideas through experiments within online mathematics assignments [12]. Using systems like E-TRIALS, students are randomized between different assignment-level interventions and complete a posttest at the end of their assignment to evaluate their learning. Experiments in E-TRIALS have shown that providing explanations, hints, or scaffolding questions to students tends improves their performance more than simply providing them with the answer after an incorrect attempt [18]. Experiments in E-TRIALS have evaluated more than just instructional intervention based experiments. For example, experiments have shown that students' learning was negatively impacted by interjecting motivational messages into their mathematics assignments [18].

Although assignment-level experiments provide some relevant information to online program designers, these designers are faced with a nearly infinite number of decisions about what features to build and how to build them. Since only one causal inference can be estimated from each manipulation [10], designing assignment-level experiments for each potential impactful variant of a feature is often infeasible. Rapid online educational experimentation provides a more efficient alternative more traditional assignment-level experiments by assigning students to a condition at each problem and instead of requiring students to complete a posttest, using the student's performance on the subsequent problem as the outcome.

One example of rapid online educational experimentation is the TeacherASSIST system, which randomizes students between crowdsourced educator generated hints and explanations. In this system, there were over 7,000 support messages produced by 11 educators. These support messages consisted of hint messages or worked explanations in both text and video form. These educator created problem-level support messages produced an average positive effect on student performance [13, 17] and more work is being done to understand the nuanced effects of each tutoring message [15]. This system has allowed for a much more efficient deployment of experiments and evaluation of feature nuances.

2.2 Unconfounded Outcomes For Rapid Online Experiments

In order for rapid online experimentation to increase the number of casual inferences made, we must identify outcomes that are unconfounded by the other experimental manipulations to which a student was exposed. Distal outcomes, such as end-of-unit or assignment-level posttest scores, do not allow a researcher to determine which of the treatments the student was exposed to during the experiment produced the effect. An alternative, used by [13, 17] to evaluate TeacherASSIST, is to use data from the problem students completed directly after the experimental condition, i.e., next problem measures.

Although individual students' behaviors and performance may be influenced by the aggregate of experimental manipulations within an assignment, the average difference in next problem measures is unconfounded due to the random assignment at the problem level. Next problem measures are unconfounded by either the prior experimental conditions or next problem experimental conditions because the assignment to each condition is independently random and therefore the effects of the prior and post-conditions are zero. Therefore, the remaining difference in the next problem measures between treatment and control is an unconfounded measure of the treatment effect.

2.3 Surrogate Measures

Although measures taken during the next problem after the experiment, such as next-problem correctness, are unconfounded by other experiments within the problem set, it is not yet known whether these measure are good estimates of distal outcomes. In assignment-level A/B testing, a researcher creates a posttest designed to measure the expected effect of the treatment condition compared to the control condition, but within online instructional interventions, the next problem was designed for pedagogical purposes, not to evaluate the effects of the intervention. Therefore, to use next problem measures to validate the impact of a condition, we must validate whether these measures assess researchers' outcomes of concern.

One way to think about these next problem measures is as surrogate measures. Surrogate measures are used in medical experiments when the outcome is either difficult to assess or distal [19]. Surrogates can either have causal or correlation relations to the outcome [11]. Validating causal surrogates requires a causal path from the treatment to the surrogate and subsequently to the outcome, such that the indirect path through the surrogate has a larger effect than the direct path through from the treatment to the outcome. Alternatively, an associative surrogate is valid when the following three criteria are met [11]:

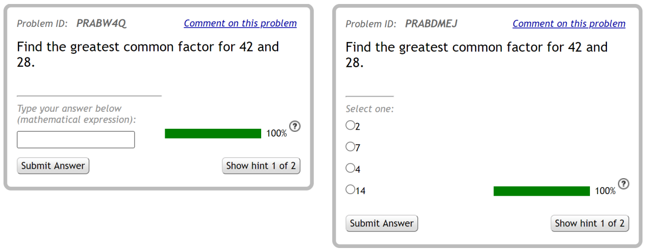
1. There is a monotonic relationship between the treatment effect on the surrogate and the treatment effect on the outcome across experiments.
2. When the treatment effect on the surrogate is zero, the treatment effect on the outcome is also zero.
3. The treatment effect on the surrogate predicts the treatment effect on the outcome.

In this work, various next problem measures are evaluated for their effectiveness as an associative surrogate measure of posttest scores.

3. DATA COLLECTION AND PREPARATION

3.1 Data Source

The data used in this work comes from ASSISTments, an online learning platform that focuses on pre-college mathematics curricula. Within ASSISTments, external researchers can run experiments at scale that compare different instructional interventions. In July, 2022 ASSISTments released a dataset of 88 randomized controlled experiments that were conducted within the platform since 2018 [18]. These experiments compared various assignment-level and problem-level interventions. For example, Fig. 1 shows the two conditions of an ASSISTments experiment in which students were randomized between receiving either open response problems, or multiple choice problems.



Find the greatest common factor for 42 and 28.

Find the greatest common factor for 42 and 28.

Figure 1: An example of two experimental conditions. In the first condition (left), students are given open response versions of mathematics problems. In the second condition (right), students are given multiple choice versions of the same problems.

In this work, only the experimental assignments from ASSISTments that had posttests were used. This ensured that any learning measures derived from a student's clickstream data on the problem immediately after receiving an intervention for the first time could be directly compared to their posttest score. A students posttest score is the fraction of problems they answered correctly on their posttest. To avoid bias from missing posttest scores, only data from experiments in which there was no statistically significant difference in students' completion rates between conditions were used, and students that did not complete the posttest were excluded from the analysis. In some contexts it would be better to impute missing posttest scores as the minimum score. However, the purpose of this work was to create a surrogate measure for posttest score in situations where it is infeasible to require students to complete a posttest, and therefore it seems more appropriate to remove missing posttest scores to ensure that the surrogate measures students' posttest scores, not their propensity to complete an assignment. This additional filtering step removed only one of the ASSISTments experiments from the analysis. Additionally, the data used in this work is limited to students who participated in the experiments prior to July 23rd, 2021. On July 23rd, 2021 all unlisted YouTube videos created prior to 2017 were made private [7]. Many of the experiments included YouTube videos uploaded prior to 2017, which were made private, ruining the experiments that contained them.

These experiments provided a rare opportunity to fairly compare next problem measures to posttest score because typically, when next problem measures are used as a dependent measure, it is because many different types of interventions are being given to a student in quick succession. However, in these experiments students are given the same intervention for each problem in the experimental assignment. Therefore, in these experiments, next problem measures measure the student's propensity to learn the material after seeing the experimental intervention for the first time, and posttest score measures the student's propensity to learn the material after seeing the same intervention multiple times, but both are evaluating the effectiveness of the same intervention. In total, 26,060 clickstream sequences of a student completing a problem and their corresponding posttest score were collected for model training and analysis across 51 different research questions within 31 different experimental assignments. These sequences and the code used to evaluate them has been made publicly available and can be found at CLICK HERE FOR BLINDED LINK.

3.2 Expert Features

As established by prior work, i.e. ([13, 17, 15]), collecting data to evaluate the effectiveness of an intervention is often limited to data from the next problem in a student's assignment, before they received another intervention. While next problem correctness was used in prior work, this work extracted four additional expert features from students' clickstream data on their next problem that have been useful predictors of student behavior in prior work [22, 23]. Table 1 describes the expert features used in this work.

3.3 Clickstream Data

In addition to expert features, this work used deep learning to create surrogate measures of learning from students' click-

Table 1: Expert Features

Feature Name	Description
Correctness	A binary indicator of whether or not the student answered the problem correctly on their first try without tutoring of any kind.
Tutoring Requested	A binary indicator of whether or not the student requested tutoring of any kind.
No Attempts Taken	A binary indicator of whether or not the student did not make any attempts to answer the problem.
Attempt Count	The number of attempts made by the student to answer the problem.
First Response Time	The natural log of the total seconds from when the problem was started to when the student submitted an answer or requested tutoring of any kind for the first time.

stream data. The clickstream data consisted of the action sequences of students within the ASSISTments tutor from the time they start the problem after they received an experimental intervention to the time they either receive another intervention or complete the problem. This short window of time is not confounded by other experimental interventions and is likely to give the clearest insight into the impact of experimental interventions being tested in quick succession.

The students' clickstream data was broken down into a series of one-hot encoded actions followed by the time since taking the last action. The first action was always "problem_started", therefore this action was dropped from students' clickstreams prior to being given to a deep learning model. The time since taking the last action was log-transformed in order to weight the difference between short time periods more than long time periods and to reduce the impact of large outliers, which are due to students walking away from their computers during assignments and returning later. Additionally, the log-transformed times are scaled within the range [0, 1]. Scaling the time within the same range as the one-hot encoded actions helps the model balance the importance of the different features. Each action sequence was equal in length to the longest action sequence, which was 12 actions. When students took less than the maximum number of actions, their action sequences were zero padded from the start of the sequence. Table 2 provides an example sequence of a student's clickstream data in which a student unsuccessfully attempted to get a problem correct twice, then took a break, then returned to their assignment, got the problem incorrect again, and then on their fourth attempt, got the problem correct. The first six columns contain all zeros because the student only took a total of six actions. This representation of students' clickstream action sequences was chosen because of previous work's success with this representation for various prediction tasks [22, 16, 23].

4. METHODOLOGY

4.1 Expert Feature-Based Models

To derive a surrogate measure of learning from the expert features, three approaches were taken. The first approach was to simply use each expert feature as a surrogate measure of learning. If an expert feature could be used as an effective surrogate measure, it would make it much easier for researchers and online learning platforms to adopt this measure, as no model fitting would be required. The second approach was to fit a linear regression on posttest score using the expert features as input. Equation 1 shows the model fit for approach two, where n is the number of students, f is the number of features, Y is an n by 1 matrix of students' posttest scores, X is an n by f matrix of students' feature values, and β is an f by 1 matrix of coefficients learned during model fitting.

$$Y = X\beta \tag{1}$$

The third approach was to fit a linear regression on the treatment effect on posttest score using the treatment effects on each expert feature as input. The third approach was included because if the goal is to predict the treatment effect on posttest score, than it might be more effective to fit a model that combines the treatment effects on different expert features into the treatment effect on posttest score than to simply predict posttest score. This would be advantageous in a scenario where there was information in the expert features that was predictive of a student's propensity to learn independent of the intervention they were given. In that scenario, a model trained to predict posttest score might learn to rely on that information, which would lead the model to predict more similar posttest scores between different experimental conditions than were actually observed. By directly predicting the treatment effect on posttest score, the model must learn to use the features that are predictive of the effect of the experimental conditions. The downside of this approach is that each research question's data is reduced to a single sample in the regression. Therefore, while the second approach had the full 26,060 samples of student data to fit on, the third approach only had 51 samples to fit on; one for each research question. Equation 2 shows the model fit for the third approach, where n is the number of students, f is the number of features, Y is an n by 1 matrix of students' posttest scores, X is an n by f matrix of students' feature values, Z is an array of conditions where 1 indicates the student was placed in the treatment condition, and 0 indicates the student was placed in the control condition, and β is an f by 1 matrix of coefficients learned during model fitting.

Table 2: A Student's Clickstream Data Sequence After Processing

Feature Name	Clickstream Data Sequence											
problem_resumed	0	0	0	0	0	0	0	0	1	0	0	0
tutoring_requested	0	0	0	0	0	0	0	0	0	0	0	0
wrong_response	0	0	0	0	0	0	1	1	0	1	0	0
correct_response	0	0	0	0	0	0	0	0	0	0	1	0
problem_finished	0	0	0	0	0	0	0	0	0	0	0	1
time_since_last_action	0.00	0.00	0.00	0.00	0.00	0.00	0.62	0.51	6.39	0.12	0.38	0.01

$$y_t = \frac{\sum_{i=1}^{n} Y_i \times Z_i}{\sum_{i=1}^{n} Z_i}, \qquad y_c = \frac{\sum_{i=1}^{n} Y_i \times (1 - Z_i)}{\sum_{i=1}^{n} 1 - Z_i}$$

$$x_t = \frac{\sum_{i=1}^{n} X_i \times Z_i}{\sum_{i=1}^{n} Z_i}, \qquad x_c = \frac{\sum_{i=1}^{n} X_i \times (1 - Z_i)}{\sum_{i=1}^{n} 1 - Z_i} \tag{2}$$

$$y_t - y_c = (x_t - x_c)\beta$$

4.2 Deep Learning Models

Two deep learning approaches were used to create a surrogate measure of learning from students' clickstream data. Both approaches trained a recurrent neural network to predict students' posttest scores given their clickstream data using Bidirectional LSTM layers [24, 6], which read the clickstream data both forward and backward to learn the relationship between students' actions and their posttest scores. Following the same intuition as the previous section, while the first model used the mean squared error of its posttest score predictions as its loss function, the second model used the squared error of the treatment effect calculated from its posttest score predictions as its loss function. Essentially, the first model was trained to predict accurate posttest scores, and the second model was trained to predict posttest scores that would lead to the same treatment effect estimates as the actual posttest scores. For context, Equation 3 formalizes the mean squared error loss function of the first approach using the same notation as Equation 4, which formalizes the custom loss function for the second approach, where Y is an array of students' posttest scores, $\hat{Y}$, is an array of predicted posttest scores, Z is an array of conditions where 1 indicates the student was placed in the treatment condition, and 0 indicates the student was placed in the control condition, n is the number of students in the array, and τ and $\hat{\tau}$ are the treatment effects of the research question calculated using posttest and the surrogate measure respectively.

$$\text{Mean Squared Error Loss} = \frac{\sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2}{n} \tag{3}$$

$$y_t = \frac{\sum_{i=1}^{n} Y_i \times Z_i}{\sum_{i=1}^{n} Z_i}, \qquad y_c = \frac{\sum_{i=1}^{n} Y_i \times (1 - Z_i)}{\sum_{i=1}^{n} 1 - Z_i}$$

$$\hat{y}_t = \frac{\sum_{i=1}^{n} \hat{Y}_i \times Z_i}{\sum_{i=1}^{n} Z_i}, \qquad \hat{y}_c = \frac{\sum_{i=1}^{n} \hat{Y}_i \times (1 - Z_i)}{\sum_{i=1}^{n} 1 - Z_i} \tag{4}$$

$$\tau = y_t - y_c, \qquad \hat{\tau} = \hat{y}_t - \hat{y}_c$$

$$\text{Treatment Effect Squared Error Loss} = (\hat{\tau} - \tau)^2$$

4.3 Model Training

To fairly evaluate the surrogate measures of learning, each model was trained and evaluated using leave-one-out cross-validation partitioned by the experimental assignment, and only the surrogate measures of learning calculated for the held out data were used to determine the surrogate measures' effectiveness. In each experimental assignment, multiple research questions are evaluated, but there is overlap in the data used to answer each of these research questions. For example, one experimental assignment evaluated the effectiveness of both video-based and text-based encouraging messages during an assignment. Both of these conditions shared the same control condition in which students did not receive encouraging messages. While there are two research questions being evaluated, if we trained a model using the data from all but one of these research questions, the data from the control condition of the held out research question would have been used to train the model. This would have given the model an unfair advantage. Therefore, when using leave-one-out cross-validation to train and evaluate the models, the data was partitioned by experimental assignment, and all the research questions in the held-out experimental assignment were evaluated using the model trained on all the other experimental assignments. This ensures that no data is shared between the training data and the held-out data.

For the expert feature-based models, an ablation study was performed to identify which combination of features lead to the highest correlation between surrogate measure and posttest treatment effects. In this ablation study, the models were trained first using all of the expert features as input, and then models were trained using all but one of the features. If any of the all-but-one-feature models out-performed the model with all the features, then that model became the best model so far, and more models were trained using all but one of the features in the new best model. Eventually, the best model will not have improved from removing any of its features, denoting that this model has the optimal set of features as input.

For the deep learning models, the models were initialized, trained, and evaluated ten times, and the average of all these evaluations was used to determine the quality of the deep learning models predictions as a surrogate measure. Unlike linear regressions, neural networks cannot be solved for the optimal value of their coefficients. Instead, a neural network's weights, which are akin to a linear regression's coefficients, are randomly initialized, and then gradient descent is used to optimize them. These random initializations can lead to more or less optimal weights at the end of training. Therefore, by training the model multiple times using different random initializations and averaging the results, the evaluation of the model's surrogate measure is more reliable.

Additionally, deep learning models are highly nonlinear and are prone to over-fitting on the data, which leads to worse predictive accuracy on the held-out data. To address this, only half of the data used for training the model were used to optimize the weights for the first approach. The other half of the data was used as a validation set. The prediction error on the validation set was calculated each time the model's weights were updated. Once the prediction error on the validation set began to increase, training was stopped, because any further reduction in prediction error on the training data would be due to over-fitting on the training data, as opposed to learning the underlying relationship between students' clickstream data and their posttest scores. For the second approach, the treatment effect loss function made it more difficult for the model to learn the relationships in the data because all predictions for a single experiment were reduced to a single loss value, making it more difficult to properly attribute blame for predictive error to the weights in the model. Therefore, none of the data was used for validation during the second approach. This provided the neural network with as much information as possible. Instead, over-fitting was prevented by training the model used in the second approach for about the same number of training steps taken by the model trained for the first approach before it began to over-fit.

4.4 Evaluation of Surrogate Measures

To reiterate from Section 2.3, a surrogate measure must meet three criteria [11]:

1. There is a monotonic relationship between the treatment effect on the surrogate and the treatment effect on the outcome across experiments.
2. When the treatment effect on the surrogate is zero, the treatment effect on the outcome is also zero.
3. The treatment effect on the surrogate predicts the treatment effect on the outcome.

Criteria 1 and 3 can be simultaneously evaluated by looking at the Pearson correlation between the treatment effect on the surrogate measures and the treatment effect on posttest score because a high Pearson correlation between two measures indicates that there is a monotonic linear relationship between them [2], and the linearity implies predictability. The higher the Pearson correlation between treatment effects across all research questions, the more effective the surrogate measure is. Using the same terminology from Equation 4, the goal is to maximize $corr(\tau, \hat{\tau})$.

To evaluate Criteria 2, after the surrogate measures were used to determine the treatment effects for the different research questions, a linear regression was fit to predict the treatment effect on posttest given the treatment effect on one of the surrogate measures and an intercept. If the coefficient of the intercept is small and statistically insignificant, then there is no evidence that Criteria 2 was violated. Therefore, the best surrogate measure was determined to be the measure with the highest Pearson correlation between its treatment effects and the posttest treatment effects across all the research questions (Criteria 1 and 3), as long as the measure did not have a significant intercept when its treatment effects were used to predict the posttest treatment effects (Criteria 2).

4.5 Experiment Analysis

It is not only important to identify the best surrogate measure, but also to understand the impact that using this measure of learning would have on analyzing A/B tests and educational experiments. Therefore, after each surrogate measure of learning was evaluated, the treatment effect on both posttest score and the best surrogate measure along with the 95% confidence interval of these treatment effects were calculated for each research question using a simple difference in means between the treatment and control groups in each research question [25]. The treatment effects on the surrogate measure were then compared to the treatment effects on posttest score.

5. RESULTS

5.1 Evaluation of Surrogate Measures

The treatment effect of each research question was calculated using each surrogate measure described in Sections 4.1 and 4.2. To evaluate whether the surrogate measures met Criteria 1 and 3 from Section 4.4, the treatment effects on each surrogate measure across all the research questions were correlated with the treatment effects on posttest score. Table 3 reports the different surrogate measures, the Pearson correlation [2] of their treatment effects, and the statistical significance of these correlations.

Of all the expert features, correctness and tutoring requested were the only two features whose treatment effects were statistically significantly correlated with the treatment effect on students' posttest scores. Correctness had a positive correlation with posttest score, indicating that students that got the next problem correct on their first try without any support tended to have higher posttest scores than those who did not, and tutoring requested had a negative correlation with posttest score, indicating that students that requested tutoring on the next problem tended to have lower posttest scores than those who did not. The direction of these correlations makes intuitive sense, as one would expect students who struggle to answer mathematics problems correctly during their assignment to have difficulty on their posttest as well.

Table 3: The Correlations between Surrogate Measure and Posttest Score Treatment Effects

Surrogate Measure	**Treatment Effect Correlation with Posttest Score**	**Correlation p-value**
Expert Features as a Surrogate Measure (Section 4.1, Approach 1)		
Correctness	**0.62**	**<0.001**
Tutoring Requested	-0.59	<0.001
No Attempts Taken	-0.01	0.935
Attempt Count	-0.16	0.264
First Response Time	0.04	0.784
Expert Features Used to Predict Posttest Score (Section 4.1, Approach 2)		
Posttest Prediction	0.62	<0.001
Expert Feature Treatment Effects Used to Predict Treatment Effect on Posttest (Section 4.1, Approach 3)		
Treatment Effect Prediction	0.50	<0.001
Deep Learning Posttest Prediction with Mean Squared Error Loss (Section 4.2, Approach 1)		
Posttest Prediction	0.60	<0.001
Deep Learning Posttest Prediction with Treatment Effect Squared Error Loss (Section 4.2, Approach 2)		
Posttest Prediction	0.49	<0.001

When performing the ablation study to identify the optimal set of expert features for the linear regression used to predict posttest score (Section 4.1, Approach 2), the highest performing model used only correctness. Interestingly, no other feature could be used in combination with correctness to improve the model's predictions. Therefore, using this linear regression to predict posttest is an equivalent surrogate measure to just using correctness as a surrogate measure itself.

When performing the ablation study to identify the optimal set of expert features for the linear regression used to predict treatment effect on posttest (Section 4.1, Approach 3), the highest performing model used tutoring requested and attempt count. Interestingly, correctness, while being the best and only feature used to predict posttest score, was not as effective at directly predicting treatment effect. Ultimately, this approach was inferior to the other approaches at identifying surrogate measures using expert features.

To evaluate Criteria 2 from Section 4.4, a linear regression was fit for each surrogate measure using data from all the research questions to predict the treatment effect on posttest given the treatment effect on the surrogate measure and an intercept. Table 4 reports the different surrogate measures, the coefficients of their linear regressions' intercepts, and and the statistical significance of these coefficients.

There was little evidence that any of the surrogate measures violated Criteria 2. Only the deep learning model with treatment effect squared error loss had an intercept coefficient that was close to statistically significant, but the p-value of 0.050 is rounded down, and that model was not a contender for best model based on the results in Table 3. Therefore, the best surrogate measure was simply next problem correctness, because the treatment effect on no other feature nor any model prediction was more correlated with the treatment effect on posttest than treatment effect on next problem correctness.

5.2 Experiment Analysis

After identifying next problem correctness as the best surrogate measure of learning, the treatment effects on posttest and on next problem correctness were calculated for each research question along with their confidence intervals. Figure 2 plots the treatment effect and confidence interval using both measures for each research question, sorted from largest to smallest posttest confidence interval. Figure 2 shows that while next problem correctness tends to lead to wider confidence intervals, it also tends to lead to larger treatment effects.

Additionally, Figure 3 shows a confusion matrix comparing the significant findings when using both measures. Only five of the 51 research questions had significant findings when using posttest score as a measure of learning. Using next problem correctness as a measure of learning resulted in six significant findings, but only one of these findings is found when using both measures to perform the analysis. However, the lack of common significant findings should not be discouraging. There is typically a sparsity of significant findings in online educational experiments, and the most important result is that the two learning measures never disagreed on which condition is better when they both identified a statistically significant difference between conditions.

6. DISCUSSION

Ultimately, next problem correctness was the best surrogate measure of learning. The treatment effect on next problem correctness had the highest Pearson correlation with the treatment effect on posttest, and there was no evidence that the treatment effect on next problem correctness was not zero when the treatment effect on posttest was zero, which satisfies all three criteria discussed in Section 2.3. It was not expected that one of the simplest of the surrogate measures, which had been used previously despite no empirical evidence to support that choice, would be the best surrogate. One possible reason for why the predictive models did not perform well is that the behavior of students within an experiment could be highly dependent on the material in the assignment. For example, geometry problems might on average take more time to answer than algebra problems, which would make students first response time less informa-

Table 4: The Correlations between Surrogate Measure and Posttest Score Treatment Effects

Surrogate Measure	**Intercept Coefficient**	**Intercept Significance p-value**
Expert Features as a Surrogate Measure (Section 4.1, Approach 1)		
Correctness	-0.0084	0.133
Tutoring Requested	-0.0059	0.293
No Attempts Taken	-0.0066	0.340
Attempt Count	-0.0080	0.293
First Response Time	-0.0073	0.177
Expert Features Used to Predict Posttest Score (Section 4.1, Approach 2)		
Posttest Prediction	-0.0085	0.131
Expert Feature Treatment Effects Used to Predict Treatment Effect on Posttest (Section 4.1, Approach 3)		
Treatment Effect Prediction	-0.0098	0.152
Deep Learning Posttest Prediction with Mean Squared Error Loss (Section 4.2, Approach 1)		
Posttest Prediction	-0.0073	0.198
Deep Learning Posttest Prediction with Treatment Effect Squared Error Loss (Section 4.2, Approach 2)		
Posttest Prediction	1.94	0.050

tive of their learning because it is in part dependent on the subject matter. Methods like Knowledge Tracing and Performance Factor Analysis, which measure students' mastery of mathematics concepts, take into account the knowledge components of the students' assignments when predicting student performance to compensate for this dependence [4, 14]. By providing the models with more nuanced information about student behavior, it is possible they were picking up on behavioral trends that were not generalizable across experiments. Additionally, the sample size of the data was fairly low. Only 51 research questions were used in this analysis, and it is likely that data from more experiments testing a greater variety of interventions would help the models learn to differentiate between generalizable trends and trends specific to subsets of experiments.

These reasons help to explain what may have caused the models to underperform, but from a different perspective, what caused next problem correctness to perform so well? It seems likely that next problem correctness was a strong surrogate because posttest score is simply a different measure of problem correctness. In other words, next problem correctness is a measure of whether the student got the problem immediately following the intervention correct, and posttest score is a measure of whether the student got a few problems ahead of the intervention correct. It makes sense that two measures that revolve around a student's propensity to answer problems correctly would correlate. This leads to the question: is correctness what matters? If the goal of education is ultimately to give students better, more fulfilling lives, then perhaps test scores are not what a surrogate should measure. There is plenty of evidence of test scores falling short when attempting to correlate them with things like college and career success. For example, studies have found that SAT scores do not explain any additional variance in college GPA for non-freshman college students after taking into account social/personality and cognitive/learning factors [8]. Additionally, these test scores can be biased against minority groups. For example, studies have found that SAT scores are more predictive of white students' college GPA than they are for Black or Hispanic students [26]. While these are important factors to consider, one could argue that these impacts are less relevant in the context of this work, where the goal is simply to use short patterns in students' behavior to analyze the difference in the impact of various problem-level interventions meant to help students learn how to correctly answer the following problems in their assignments. However, one should always be cognizant of the potential bias a surrogate measure could introduce.

When using next problem correctness and posttest scores to analyze the results of the 51 research questions, only six and five of the 51 research questions had significant differences between conditions respectively, but only one of these significant findings was identified by both measures. While it would be better if the two measures found more similar significant findings, as long as the two measures do not disagree on which condition is most effective when they both find something statistically significant, then there is no concern that using next problem correctness could lead a researcher to the wrong conclusion. Next problem correctness, on average, had wider confidence intervals than posttest score, but also had larger treatment effects. This may be explained simply by the more extreme nature of the next problem correctness values. To gain some intuition on why this might be the case, consider that posttest is essentially the average of many next problem correctness measures. If we think of whether a student gets a problem correct as a random variable, then one can see how the average of many random variables will tend to be closer to the expected value than a single random sample. The variance of students' posttest scores can therefore be expected to be lower than the variance of students' next problem correctness, which would cause the confidence interval of the treatment effect on posttest to be smaller as well.

6.1 Limitations and Future Work

While in this work next problem correctness was found the be the best proximal surrogate measure for posttest score, there are some factors that could limit the generalizability of these findings. Firstly, this work uses data entirely from ASSISTments Skill Builder assignments. In these assignments, students are given a series of mathematics problems on the same skill, and are given immediate feedback on each prob-

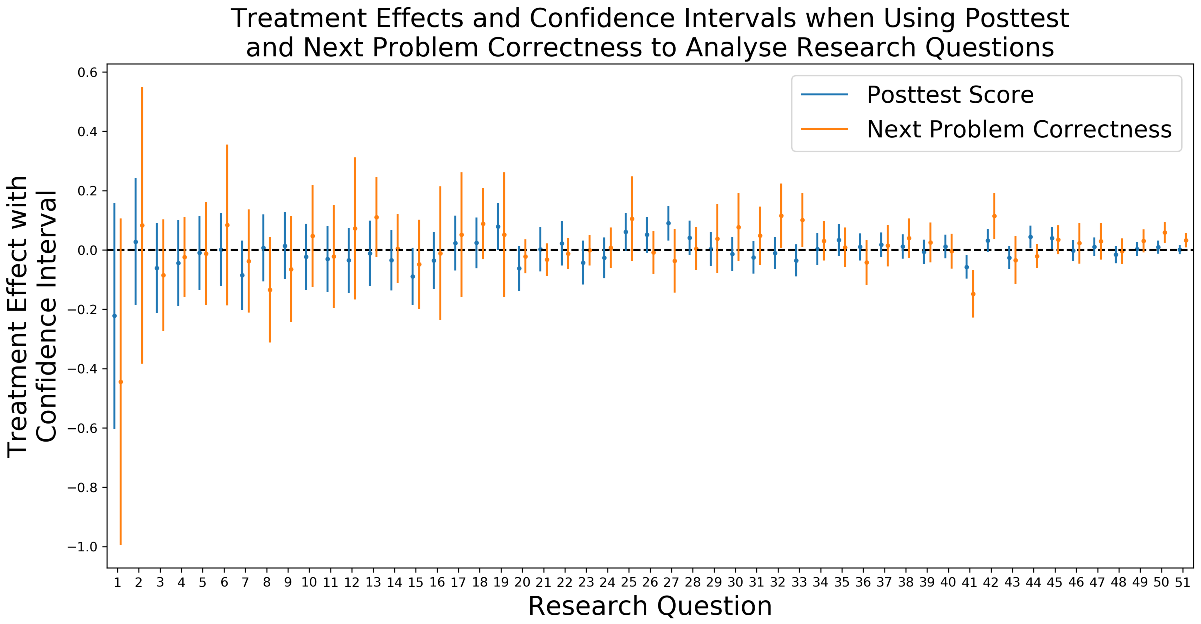


Figure 2: A plot of the treatment effect and confidence interval determined using posttest score and next problem correctness for each research question.

lem as they complete it. Next problem correctness could be especially relevant in this context because the next problem is guaranteed to evaluate the same knowledge components as the previous problem. In assignments where interleaving [3] is used, the problem following an intervention could be only tangentially related to the problem for which the intervention was provided, and thus a student's performance on the next problem would not be a good measure of the effectiveness of the intervention. In the future, using next problem correctness as a surrogate measure should be evaluated in other kinds of online learning environments, perhaps in contexts where the content students see is chosen adaptively. In this scenario, students will see different problems following an intervention, and combining the next problem correctness of multiple problems could have positive or negative effects on next problem correctness's value as a surrogate measure of learning.

Additionally, in this work, only 51 different research questions were used to evaluate the quality of different measures, with a total of 26,060 samples. It is possible that some of the model based attempts at creating a surrogate measure of learning would be more successful if given more data from a wider variety of situations in which A/B testing was performed. Having a larger and more diverse dataset to train the models from also opens up the possibility to train multiple specific models for different subgroups of users or experiments. With the limited data in this work, it was unlikely that splitting the data into subgroups would have helped any of the models. However, with more data it could be the case that a model trained on students from a specific socioeconomic background would be more effective at interpreting behaviors specific to those students. It could also be the case that training a model for a specific type of experiment, for example, experiments that alter the way in which students must answer the question as opposed to experiments that alter the support messages students receive, could improve the model's ability to pick up on different student behaviors associated with these different experiments. In the future, if more data becomes available, models trained on subgroups should be explored.

7. CONCLUSION

In this work, we attempted to derive and validate an effective surrogate measure of learning for use in online learning platforms where rapid A/B testing is used to compare problem-level instructional interventions at scale. To accomplish this, a variety of proximal surrogate measures for posttest score were created through feature engineering, regression, and deep learning. After evaluating each surrogate measure by ensuring it met the criteria for an associative surrogate as described in [11], students' next problem correctness was determined to be the best surrogate. When comparing the treatment effect on posttest score to the treatment effect on next problem correctness across 51 different research questions, both measures determined that approximately 10% of the research questions had statistically significant treatment effects, but both of the measures shared only one statistically significant finding. Although there was not much overlap in these significant findings, both measures agreed on which condition was most effective when they both found a significant treatment effect. Additionally, using next problem correctness as a measure lead to larger treatment effects with wider confidence intervals than using posttest score.

Follow-up work should be done to validate next problem correctness as a measure of learning in different domains and for different learning environments. Moving forward, using next problem correctness as a measure of learning within online learning platforms could be an effective way to evaluate stu-

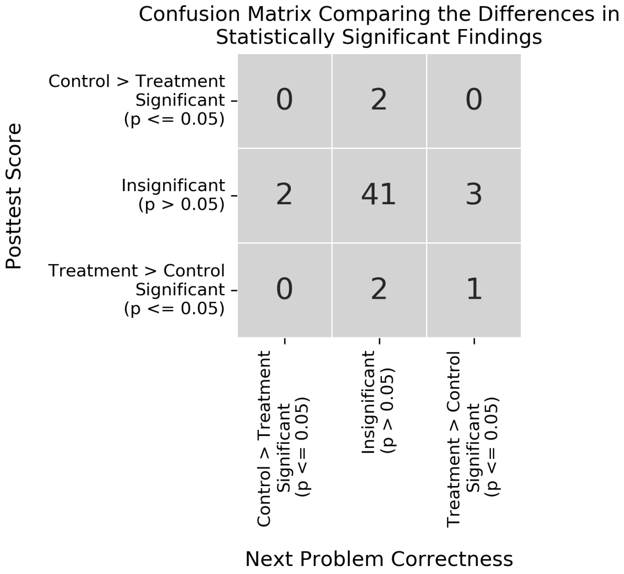


Figure 3: A confusion matrix comparing the differences in statistically significant findings when using posttest score and next problem correctness as measures of learning.

dents' progress and compare problem-level interventions to each other. We hope this work can help support the learning analytics community by providing a way to rapidly evaluate new instructional methods and interventions.

8. ACKNOWLEDGEMENTS

Blinded for review.

9. REFERENCES

[1] R. S. Baker, N. Nasiar, W. Gong, and C. Porter. The impacts of learning analytics and a/b testing research: a case study in differential scientometrics. *International Journal of STEM Education*, 9(1):1–10, 2022.

[2] J. Benesty, J. Chen, Y. Huang, and I. Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer, 2009.

[3] S. K. Carpenter. Spacing and interleaving of study and practice. *Applying the science of learning in education: Infusing psychological science into the curriculum*, pages 131–141, 2014.

[4] A. T. Corbett and J. R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4(4):253–278, 1994.

[5] C. for Research and J. H. U. Reform in Education. Evidence for essa, 2022.

[6] F. A. Gers, J. Schmidhuber, and F. Cummins. Learning to forget: Continual prediction with lstm. *Neural computation*, 12(10):2451–2471, 2000.

[7] Google. Older unlisted content, 2022.

[8] B. Hannon. Predicting college success: The relative contributions of five social/personality factors, five cognitive/learning factors, and sat scores. *Journal of Education and Training Studies*, 2(4):46, 2014.

[9] N. T. Heffernan and C. L. Heffernan. The assistments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, 24(4):470–497, 2014.

[10] G. W. Imbens and D. B. Rubin. Rubin causal model. In *Microeconometrics*, pages 229–241. Springer, 2010.

[11] . G. T. Joffe, M. M. Related causal frameworks for surrogate outcomes. *Biometrics*, 65(2):530–538, 2009.

[12] K. S. Ostrow, D. Selent, Y. Wang, E. G. Van Inwegen, N. T. Heffernan, and J. J. Williams. The assessment of learning infrastructure (ali) the theory, practice, and scalability of automated assessment. In *Proceedings of the sixth international conference on learning analytics & knowledge*, pages 279–288, 2016.

[13] T. Patikorn and N. T. Heffernan. Effectiveness of crowd-sourcing on-demand assistance from teachers in online learning platforms. In *Proceedings of the Seventh ACM Conference on Learning@ Scale*, pages 115–124, 2020.

[14] P. I. Pavlik Jr, H. Cen, and K. R. Koedinger. Performance factors analysis–a new alternative to knowledge tracing. *Online Submission*, 2009.

[15] E. Prihar, A. Haim, A. Sales, and N. Heffernan. Automatic interpretable personalized learning. In *Proceedings of the Ninth ACM Conference on Learning@ Scale*, pages 1–11, 2022.

[16] E. Prihar, A. Moore, and N. Heffernan. Identifying struggling students by comparing online tutor clickstreams. In *International Conference on Artificial Intelligence in Education*, pages 290–295. Springer, 2021.

[17] E. Prihar, T. Patikorn, A. Botelho, A. Sales, and N. Heffernan. Toward personalizing students' education with crowdsourced tutoring. In *Proceedings of the Eighth ACM Conference on Learning@ Scale*, pages 37–45, 2021.

[18] E. Prihar, M. Syed, K. Ostrow, S. Shaw, A. Sales, and N. Heffernan. Exploring common trends in online educational experiments. 2022.

[19] P. R. L. Surrogate endpoints in clinical trials: definition and operational criteria. *Stat Med*, 1989.

[20] J. Renz, D. Hoffmann, T. Staubitz, and C. Meinel. Using a/b testing in mooc environments. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, pages 304–313, 2016.

[21] S. Ritter, A. Murphy, S. E. Fancsali, V. Fitkariwala, N. Patel, and J. D. Lomas. Upgrade: an open source tool to support a/b testing in educational software. In *Proceedings of the First Workshop on Educational A/B Testing at Scale (at Learning@ Scale 2020)*, 2020.

[22] A. Sales, A. Botelho, T. Patikorn, and N. T. Heffernan. Using big data to sharpen design-based inference in a/b tests. In *Proceedings of the Eleventh International Conference on Educational Data Mining*, 2018.

[23] A. C. Sales, E. Prihar, J. Gagnon-Bartsch, A. Gurung, and N. T. Heffernan. More powerful a/b testing using auxiliary data and deep learning. In *International Conference on Artificial Intelligence in Education*, pages 524–527. Springer, 2022.

[24] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

[25] B. L. Welch. The generalization of ‘student’s’problem when several different population varlances are involved. *Biometrika*, 34(1-2):28–35, 1947.

[26] R. Zwick and J. C. Sklar. Predicting college grades and degree completion using high school grades and sat scores: The role of student ethnicity and first language. *American Educational Research Journal*, 42(3):439–464, 2005.