Article

How should I teach from this month onwards? A State-Space Model that Helps Drive Whole Classes to Achieve end-of-year National Standardized Test Learning Targets

Obed Ulloa [1] and Roberto Araya [2]*

[1] Centro de Investigación Avanzada en Educación, Instituto de Educación, Universidad de Chile, Santiago 8320000; oulloa@dim.uchile.cl
[2] Centro de Investigación Avanzada en Educación, Instituto de Educación, Universidad de Chile, Santiago 8320000; roberto.araya.schulz@gmail.com
* Correspondence: roberto.araya.schulz@gmail.com; Tel.: +56-9-9599-0251 (R.A.)

Abstract: Every month teachers face the dilemma of what exercises should their students practice, and what their consequences are on long-term learning. Since teachers prefer to pose their own exercises, this generates a large number of questions, each one attempted by a small number of students. Thus, we couldn't use models based on big data such as deep learning. Instead, we developed a simple to understand state-space model that predicts end-of-year national test scores. We used 2,386 online fourth-grade mathematic questions designed by teachers and each attempted by some of the 500 students in 24 low socioeconomic schools. We found that the state-space model predictions improved month-by-month and that in most months it outperformed linear regression models. Moreover, the state-space estimator provides for each month a direct mechanism to simulate different practice strategies and compute their impact on the end-of-year standardized national test. We built iso-impact curves based on two critical variables: the number of questions solved correctly in the first attempt and the total number exercises attempted. This allows the teacher to visualize the trade-off between asking students to do exercises more carefully or doing more exercises. To the best of our knowledge, this model is the first of its kind in education. It is a novel tool that supports teachers drive whole classes to achieve long-term learning targets.

Keywords: Digital Systems; Educational Systems; State-space Models; Optimal Control; Long-term learning prediction; Learning Analytics

check for updates

Citation: Ulloa, O.; Araya, R. How should I teach from this month onwards? A State-Space Model that Helps Drive Whole Classes to Achieve end-of-year National Standardized Test Learning Targets. *Preprints* **2022**, *10*, 0. https://doi.org/10.3390/xxxxx

Academic Editor: Firstname Lastname

1. Introduction

Teachers have to constantly estimate the probable effect of their decisions on the long-term learning achieved by their students. For each learning objective of the national curriculum, teachers have to make the students practice enough exercises. A key decision they must make is between asking students to do exercises more carefully or doing more exercises. Thus, every month teachers face the dilemma of what and how many exercises should their students practice, and what their consequences are on long-term learning. For that they need to have a good prediction mechanism. This dilemma presents several challenges.

First, the mechanism should predict their students' long-term learning. This is not straightforward because long-term learning is not always the same as short-term learning [27]. This is a well-known unintuitive phenomenon, where fast progress measured on real-time generates an illusion of mastery in the students [11]. This illusion makes it difficult for the teachers to make a good prediction of the long-term learning of each of their students.

Second, we must consider that teachers prefer to design their own questions. This generates a huge diversity and number of questions, each answered by very few students. Therefore, the data gathered is not suitable for typical big-data mining such as deep learning algorithms. For example, [14] found that logistic regression leads on data-sets of moderate size, whereas Deep Knowledge Tracing leads on data-sets of large size.

 doi:10.20944/preprints202208.0124.v1

Third, from one month to the next, teachers need a simple update mechanism. For example, they could use the average score on the answers to questions posed on the current month and an accumulative score of the previous months. They should no need the whole history of scores of all answers to previous questions. The mechanism must also combine in a simple way, hopefully additively, with information on other personal attributes of students, of their class and their school. These are personal attributes such as student beliefs, course attributes such as classroom climate, and school attributes such as its historical performance on national standardized tests.

Fourth, the mechanism should facilitate simulating different practice strategies for the coming months. These strategies together with the assessment in the current month, should be enough to predict the effect on the long-term learning. Teachers normally use grade point average (GPA) for this purpose. GPAs are very simple to understand. They are just averages of partial grades obtained across the year. Teachers could welcome a better predictor of long-term learning if it is as simple and interpretable as GPAs. If we were to use nonlinear models for this prediction, such as gradient boosted machines, k-nearest neighbors or random forests, this goal would not be possible. Non-linear intertwined relationship between the estimated model parameters and the predicted score are difficult to interpret. Most Machine Learning models are designed with features and combinations of them that maximize prediction performance. There is no consideration to generate solutions that are meaningful to domain experts, such as teachers [33]. They are typically black boxes that are not transparent to users [10]. However, interpretability is critical for teachers to adopt the model, to simulate instructional strategies to follow in the coming months, and to use them to base their decisions.

Therefore, in order to improve the quality of education, teachers need to be able to predict for each student her long-term learning and how different strategies can impact this prediction. One way to do so, is to know the current state of long-term learning of each student, and how different strategies could interact with each student's current state. Thus, based on an estimation of the actual state of long-term learning of each student and the effects of different strategies that actuate on the current state of each student, the teacher has to adjust her strategies of exercise practice in order to reach a specified long-term learning target.

1.1. Related work

[1] developed linear predictors of students' scores of end-of-year state test from dynamic testing metrics developed from monthly intelligent tutoring system data. They analyzed the data logs of 362 students, even though only 105 students had complete data in each of the months. They found that logs from an online tutoring system provide better end-of-year predictions than using paper and pencil benchmark tests. They found that R adjusted is 0.637. However, all students attempted to solve a similar set of items. One of the challenges is that teachers prefer to have the flexibility to adapt the exercises to their own experience. The authors didn't consider the situation where teachers select or design their own set of exercises. [1] also compared month-by-month their linear predictors without pretest with baselines. They found an increasing performance as they gather process information of more months. On a sample of 23,000 learners in Grades 6, 7, and 8 over three academic years, [32] analyzes the relative contribution of different types of learner data in statistical models that predict scores on summative assessments. They use six different categories of statistical models that predict summative scores. One of the best model categories turned out to be Stepwise Linear Regression (SLR). In the best year, it achieved a R^2 of 0.734.

[19] analyze 62 papers published between 2010 to 2020 that predict mainly university students' academic outcomes and in computer sciences courses. They focus on the forms in which the learning outcomes are predicted, the predictive analytics models developed to forecast student learning, and the dominant factors impacting student outcomes. The authors concluded that the best performing predictive models were the Hybrid Random

Forest, Feedforward 3-L Neural Network, and Naive Bayes. However, there is a lack of explanations that go beyond predicting the student performance. [24] argues that multivariate linear regression models are more adequate to analyze which variables have a positive or negative impact on the predicted learning gain than deep learning nonlinear models. The authors claim that these models don't have the best performance in terms of metrics such RMSE or R^2, but they provide more understanding. We claim that state space models provide an even better understandability.

[17] propose a state-space model of skill formation in a home visiting program implemented in rural China. It is a dynamic learning model for multiple skills that includes the sources of learning in the early life-cycle. They use the model to estimate the impact of different interventions. An interesting feature of the model is that they extend the traditional measure of accuracy (% of correct answers) in two new variables: time to mastery and how stable the child's performance is after first mastery. The authors also include different rates of learning among students.

1.2. Research questions

In this paper we have three research questions.

Research Question 1: To what extent can teacher designed questions (low or no stake quizzes) help predict students' long term learning as measured by end-of-year standardized state tests?

Research Question 2: To what extent a Markovian hidden state of the student accumulated knowledge up to the current month can be estimated so that along with the probable deliberate practice on the next months are sufficient enough to predict with good accuracy the end-of-year standardized state tests?

Research Question 3: To what extent a state-space model can help teachers to visualize the trade-off between asking students to do exercises more carefully or doing more exercises, and thus help drive whole classes to achieve long-term learning targets?

This hidden state would capture the knowledge stored in long-term memory as well as the ability to retrieve it and apply it properly.

The objective is to optimize long-term learning as measured in end-of-year state standardized tests. This is different to short-term learning, and particularly to the performance on the next question. One key restriction of the problem is that the questions of the formative assessments have been proposed by the teacher. Therefore, they are questions that have not being previously piloted. Moreover, for every question there are very few students that have attempted it. This situation makes it very different of the ones where big data algorithms can be used.

2. Materials and Methods

In this paper, we use the questions and answers in ConectaIdeas [2–6,8]. This is an online platform where students answer closed and open-ended questions. On ConectaIdeas, teachers develop their own questions, designing them from scratch, or they select them from a library of questions designed previously by other teachers. Then teachers use those questions to build their own formative assessments composed by sets of 20 to 30 questions. Students answer them in laboratory sessions held once or twice a week, or at home.

The decision to work with fourth grade students is due to three facts. First, every year in Chile there is a national standardized end-of-year test at that grade level. This is only true for 4th, 8th and 10th grade. This fact is therefore very important for schools and teachers. Second, at a younger age the return on investment is greater than with older students, as shown by the Heckman curve [16], see however [23]. Third, at fourth grade students can autonomously login and answer on their devices.

We use data from 24 fourth-grade courses from 24 different schools with low socio-economic status in Santiago, Chile. From those classes there are a total of 256 boys and 244 girls that took the pretest and the national standardized end-of-year test. On average, we have 20.8 students per class. The average age at the beginning of the school year was 9.5

years with SD of 0.6 years. The courses belong to high vulnerability schools. The official vulnerability index of the schools is 0.28 SD above the national average. The historical performance of these schools in mathematics is very low. The historical score on the national standardized end-of-year test of the previous 3 years of the schools is 0.74 SD below the national average. All data corresponds to the 2017 school year.

Unlike intelligent tutoring systems, teachers select a list of questions for the whole class and even writes their own questions. Those students who manage to make 10 correct questions in a row in the session become candidates to be tutors of the session. From that list the teacher selects 2 or 3 students as tutors. So, when a student asks for help, a tutor can help him. In those moments of the session when there is an option to select a peer or the teacher, students prefer a student tutor [9]. Additionally, the ConectaIdeas early warning system tracks the behavior of students' responses. If a student makes too many mistakes or is answering too slow, the system alerts the teacher and then she helps the student or assigns a tutor to help the student.

From a total of 2,386 closed questions, only 60 questions were attempted by at least 90% of the students. Moreover, only 8.6% of the closed questions were attempted by at least one student of all of the participating classes. On the other hand, from a total of 1,071 open-ended questions, only 8 questions were answered by more than 5% of the students. In this paper we do not analyze the answers to open-ended questions. Although there are much fewer open questions than closed ones, the answers to open-ended questions also manage to make a contribution to learning and improve the prediction of the score in the national standardized end-of-year test [31]. This improvement is achieved even when the standardized national test at the end-of-the-year only has closed questions and they only are of the multiple-choice type.

Each month the students carried out closed-ended exercises on the ConectaIdeas platform. The exercises are from the five strands of the fourth-grade Chilean mathematics curriculum:

1. Numbers and operations
2. Patterns and Algebra
3. Geometry
4. Measurement
5. Data and probabilities

Table 1 shows the average number of closed-ended exercises performed each month and the corresponding standard deviations. The table shows that the students performed an average of 1089 exercises during the year. The month with the most exercises performed was October, just before the year-end national test. The strand with the most exercises is Numbers and Operations. 65.2% of the exercises performed are from that strand.

Table 1. Number of closed ended exercises performed by month for each strand, and monthly average.

Strand	March	Apr	May	June	July	Aug	Sept	Oct	Average
Numbers	6	120	123	69	56	122	119	93	89
Patterns	0	2	4	18	7	1	14	27	9
Geometry	0	0	11	22	7	24	17	64	18
Measurement	0	0	5	35	11	17	11	28	13
Data	0	1	2	3	5	3	7	30	6
Total	6	124	146	147	86	168	168	243	136

In order to select the features to be used in the models, we calculated the correlation of 68 variables with the national standardized end-of-year test. From those we chose 28 variables with which we trained linear models in a set of training courses to predict the end-of-year scores in the national standardized end-of-year test. We then selected the variables that achieved the best prediction in a set of test courses. These variables were the pre-test, a variable of students' beliefs (the degree of agreement on a scale of 1 to 5 with "Mathematics is easy for me"), one variable that indicates the climate of the class (the average of all the students of the class to the degree of agreement of each student on a scale of 1 to 5 with "My behavior is a problem for the math teacher"), and two performance variables on the formative assessments. These performance variables are the percentage of correct answers on the first attempt, and the difference between the number of questions answered correctly on the first attempt and the number of questions answered correctly not on the first attempt. Every time the student makes a mistake, the system notifies him of the error. Thus, in the long run the student correctly answers the question, and thus can continue with the next question. We normalize each of these variables.

2.1. The sequence of Linear Regression models

We have studied before both linear regression models [2,7,8,29,30], and state space models [29,30]. In this paper we develop them in much more detail.

The first type of models is a sequence of linear regression models. For each month of the school year, we trained an optimal model. The model for each month is a Linear Regression for the Month with an Optimal Forgetting Rate (LRMOFR). This is a model that for each month finds an optimal combination of pre-test, one student and one classroom feature, and the two historic performance variables up to the current month on formative assessments. The historic performance is the addition of all the discounted average performance of each month up to the current month. The discounting is computed with an optimal forgetting rate. Thus, for each month we have a prediction model. The model for a new month does not use the finding of the optimal parameters of the previous months. They are completely new optimization models. The only common optimal parameters are the forgetting or discount rate and the regression intercept.

The central problem in this type of models is to find the contribution to the prediction of each component of the model. It is then necessary to find in each month how much an increase in one standard deviation of the pretest contributes to the national standardized end-of-year test (SIMCE) test score, how much an increase in one standard deviation in the student's subjective appreciation that math is easy contributes to the SIMCE score, how much a one standard deviation increase in poor classroom climate contributes to the SIMCE score, and how much a one standard deviation increase in both historical performance features contributes to the SIMCE score. Here, the historical performance until each month is the weighted average of the monthly percentage of exercises solved correctly on the first attempt in the formative evaluations from the beginning until the month and the difference between the number of exercised solve correctly in the first attempt and the number of exercises solved correctly not in the first attempt.

More precisely, for a student i from a school j, with information until month t , the Linear Regression for each Month with Optimal Forgetting Rate Model (LRMOFR) predicts the SIMCE score:

$$\widehat{y}_{i,j,\tau} = \gamma_\tau p_i + \beta_0 + \beta_{1,\tau}\Theta^{(1)}_{i,j} + \beta_{2,\tau}\Theta^{(2)}_{i,j} + \frac{\sum_{t=1}^{\tau}\rho^{\tau-t}\left(\sum_{k=1}^{5}\delta_{k,\tau}a_{i,j,k,t}+\eta_{k,\tau}d_{i,j,k,t}\right)}{\sum_{t=1}^{\tau}\rho^{\tau-t}} \tag{1}$$

- p_i: pretest of the student i ($p_i \in \mathbb{R}$, normalized).
- $\Theta^{(1)}_{i,j}$: class j average for "My behavior is a problem for the math teacher".
- $\Theta^{(2)}_{i,j}$: student i response to "Math is easy for me".
- $a_{i,j,k,t}$: percentage of exercises solved on the first attempt for a student i, on the k strand, at month t (normalized for each strand-month).
- $d_{i,j,k,t}$: the difference between exercises solved on the first attempt and those that took more than one for a student i, on the k strand, at month t (normalized for each strand-month).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2022 doi:10.20944/preprints202208.0124.v1

On the model, the parameters are $\rho, \beta_0, \beta_{1,\tau}, \beta_{2,\tau}, \delta_{k,\tau}, \eta_{k,\tau}$ and γ_t where $\tau \in \{1, \ldots, 8\}$ (from March until October since the test was on November) , $k \in \{1, \ldots, 5\}$. This means that the ρ and β_0 parameters are shared by every $\widehat{y}_\tau$ model and each has it's specific $\beta_{1,\tau}, \beta_{2,\tau}, \delta_{k,\tau}, \eta_{k,\tau}$ and γ_τ estimates. Then, that makes it that each $\widehat{y}_\tau$ LRMOFR has 15 parameters and for the 8 months overall that is 106 estimates.

To find these values, we minimize the average RMSE across months:

$$
(P)\quad
\begin{array}{lrcl}
\min\limits_{\rho,\beta_0} & \multicolumn{3}{l}{\frac{1}{8}\sum\limits_{\tau=1}^{8}\sqrt{e(\rho,\beta_0,\theta_\tau)^2}} \\
\text{s.a:} & \rho & \in & [0,1] \\
 & \beta_0 & \in & [210,275] \\
 & \theta_\tau & \in & \arg\min\limits_{\tilde{\theta_\tau}}(P_\tau)
\end{array}
\qquad
(P_\tau)\quad
\begin{array}{lrcl}
\min\limits_{\theta_\tau} & \multicolumn{3}{l}{e(\rho,\beta_0,\theta_\tau) = \sqrt{\frac{1}{N}\sum\limits_{i=1}^{N}\left(y_i - \widehat{y}_{i,j,\tau}\right)^2}} \\
\text{s.a:} & \theta_\tau & = & (\gamma_\tau,\beta_{1,\tau},\beta_{2,\tau},\delta_{1,\tau},\ldots,\delta_{5,\tau},\eta_{1,\tau},\ldots,\eta_{5,\tau}) \\
 & \widehat{y}_{i,j,\tau} & = & \gamma_\tau p_i+\beta_0+\beta_{1,\tau}\Theta^{(1)}_{i,j}+\beta_{2,\tau}\Theta^{(2)}_{i,j} \\
 & & & +\dfrac{\sum\limits_{t=1}^{\tau}\rho^{\tau-t}\left(\sum\limits_{k=1}^{5}\delta_{k,\tau}a_{i,j,k,t}+\eta_{k,\tau}d_{i,j,k,t}\right)}{\sum\limits_{t=1}^{\tau}\rho^{\tau-t}}
\end{array}
$$

and (P_τ) is the OLS problem for each linear regression. Making it a 106 parameters model.

A good characteristic of these linear models is that each of their parameters is clearly understood and they have direct implications in the prediction. Their effects are independent of each other and the total effect is just an addition of each separate effect. However, it is one model per month. In each month there are 13 parameters. That means 104 parameters. If we add the 2 overall parameters, we have a total of 106 parameters. This is a huge number of parameters. It is very difficult to memorize, which can make analysis difficult for a teacher.

[30] analyzed other possibilities of models with linear regressions. The one presented in this paper has the best prediction error and at the same time the one with the least number of parameters.

2.2. *The State-Space model*

The second type of model is an accumulator model. In the first month is fed with three inputs: the contribution of the pretest, the contribution of the student's subjective appreciation that math is easy, and the contribution of poor classroom climate indicator. More precisely, as in the linear regression models, the contribution to the prediction of the national standardized end-of-year test of each of the 3 inputs is proportional to the difference in standard deviations of the input and its population average. Then on each of the next months, we add the contributions of the corresponding two monthly formative assessments performances. More precisely, from a month t to the next month $t+1$, we have:

$$
\begin{array}{rcl}
x(t+1) & = & Ax(t) + Bu(t) + w(t) \\
y(t) & = & x(t) + v(t)
\end{array}
\tag{2}
$$

where $t \in \{0, \ldots, T\}$,$x(t), y(t) \in \mathbb{R}$, $u(t) \in \mathbb{R}^{13}$, $A \in \mathbb{R}$, $B \in \mathbb{R}^{13}$ $w(t) \sim \mathcal{N}(0, Q)$ y $v(t) \sim \mathcal{N}(0, R)$ are the noise with variances $Q, R \in \mathbb{R}$.

- $x(t)$ is the hidden state that represents the long-term knowledge of a student.
- $y(t)$ are the observed measurement as a grade or score.
- $u(t)$ has historical variables at $t = 0$ and dynamic ones for $t > 0$.

With this structure, we can predict with information till a month τ with the equations:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2022 doi:10.20944/preprints202208.0124.v1

$$
\begin{array}{lcll}
\widehat{x}(\tau, t+1) & = & (A-K)\widehat{x}(\tau, t) + Bu(t) + Ky(t) & t \in \{0, \ldots, \tau\} \\
\widehat{\widehat{x}}(\tau, t+1) & = & A\widehat{\widehat{x}}(\tau, t) + B\widehat{u}(\tau, t) & t \in \{\tau+1, \ldots, T\} \\
\widehat{\widehat{x}}(\tau, T) & = & \widehat{x}(\tau, T) & \\
\widehat{y}(\tau, T) & = & \widehat{\widehat{x}}(\tau, T) & \\
\widehat{x}(0) \text{ estimated} & & &
\end{array}
$$

where $\widehat{u}(\tau, t)$ is estimated repeating the last known value of $u(t)$ ($\widehat{u}(\tau, t) = u(\tau), \forall t > \tau$) and K is Kalman gain matrix (from the respective steady state Riccati equation).

This is a much simpler model than the linear regression model. The main reasons are the following five.

First it is just one model, not a sequence of models. Therefore, it requires much smaller number of parameters. The same parameters are used on the different months. We have a total of 17 parameters. This is a much smaller number than the 106 parameters in the LRMOFR models.

Second, the model is a simple accumulator. It is similar to a tank fed with water. From one moment to the next we add water to the already accumulated water. In the accumulator model, instead of water, a hidden state accumulates 10 different contributions. These are two contributions for each one of the five strands of the curriculum. Additionally, the update of the hidden state is a very simple additive mechanism. In the next month we have to just add the contributions of this new month. It is a Markovian model.

Third, to make predictions from the accumulator we need to simulate teaching strategies in the next months with an estimated performance of the student. We can input different strategies and student performance, and compute the prediction. This is a powerful tool that gives great flexibility to the teacher. In the sequence of linear regression models, the process is more complex. There is no single place to input and simulate different teaching strategies. The model is fixed and predicts assuming the teacher follows the pattern of strategies used in 2017.

Fourth, the accumulator model is more robust than the sequence of linear regression of models. The sequence of linear regression models is more prone to capture patterns that could have been produced in a certain particular month of 2017. For example, the impact on some months of a teacher strike, social riots, or the educational effect of a big natural disaster as an earthquake. On the contrary, the accumulator model has a mechanism for updating from one month to the next, and therefore fits patterns of the whole year without overfitting to each month.

Fifth, the accumulator model provides a simple conceptual construct that is similar to a grade point average that all teacher use. It is a single number that represents the long-term learning of the student. It also provides an understandable and transparent updating mechanism. Moreover, it integrates personal and historic information of the student and her classroom with the student performance on the formative assessments. It converts all the information in a single currency, and provides a change mechanism for converting different educational, personal, and social information into that currency.

In both types of models, a detailed specification of the contribution of the formative assessments is critical. Given that students respond to the formative assessments in each session and that the teacher has their performances for each month, she needs that information to do what-if type of analysis. For example, to estimate the effect on the SIMCE predictor if in a given month the student increases her performance by one standard deviation in the formatives assessment of that month. Clearly, the effect must be null in the months prior to the month in which performance increases. But in that and subsequent months, an effect should be provoked. In the language of dynamical systems, we are estimating the impulse response. That is, if in a month s the student experiences a shock or impulse of additional performance, then we estimate its effect $h(s, t)$ on the predictor of the SIMCE score that is computed in a month t.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2022 doi:10.20944/preprints202208.0124.v1

3. Results

We tried 12 linear regression models and 12 state-space models. We tested four variants corresponding to student's pretest: no pretest, with pretest, with third grade GPA instead of pretest, and with fourth grade GPA instead of pretest. On the other hand, we included three variants for performance month by month: without performance variables, for each of the 5 math strands of the curriculum we included only the percentage of correct answers in the first attempt, and for each of the 5 math strands of the curriculum we included the percentage of correct answers in the first attempt and the difference between the number of correct answers on the first attempt and the number of correct answers not on the first attempt.

In total there are 12 linear regression models and 12 state space models. We performed 200 cross validations, each time training in 18 classes and testing in 6 classes. The models with lowest RMSE were the linear regression and the space-state models with the following options: pretest and both monthly performance measures. In Table 2 we show, month by month, the percentage of times in which the linear regression model and the state-space model with those options had the lowest RMSE within the 24 models. The RMSE was measured in the testing classes.

Table 2. Percentage of times in which the model had the lowest RMSE within the 24 models in the testing classes.

Model	March	Apr	May	June	July	Aug	Sept	Oct	Average
Linear regression	20.0	19.0	4.5	4.0	1.0	5.5	28.0	40.5	15.3
Sate-Space	19.5	24.0	61.0	62.5	69.0	69.0	35.5	32.0	46.6

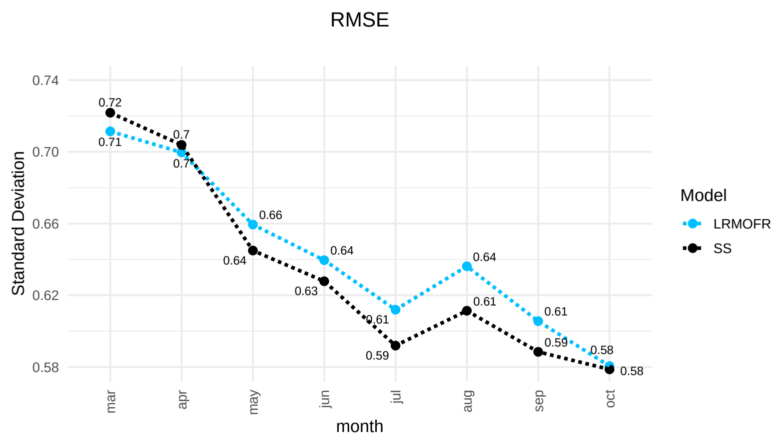
Therefore, in what follows we report only the two best models. These are the Linear Regression for each Month with Optimal Forgetting Rate (LRMOFR) and the State Space (SS) model with the following variables: pretest, the personal belief variable, the class climate variable, and the two monthly performance variables. These performance variables are the percentage of correct answers on the first attempt, and the difference between the number of correct answers on the first attempt and the number of correct answers not on the first attempt.

3.1. Prediction errors of both models

We found that in both models, their prediction error generally decrease as the school year progresses (figure 1). Coincidentally, the fit between the prediction and the SIMCE results improve as the year progresses (figure 2). This trend is expected since the models are receiving the results of the deliberate practice of the students. Every week students carry out one or two sessions and in each of them they answer about 24 questions that are part of the formative assessments. However, RMSE does not reach to zero. This is partly because the error of the SIMCE test is between 0.26 and 0.35 standard deviations. Therefore, it is not possible for the models to go below that level of error.

It is interesting to observe that in August both indicators worsen a little bit. Apparently, this is due to the two-week winter break in July. The models receive much less information. In July, students take half of the formative assessments of what they do on the other months.

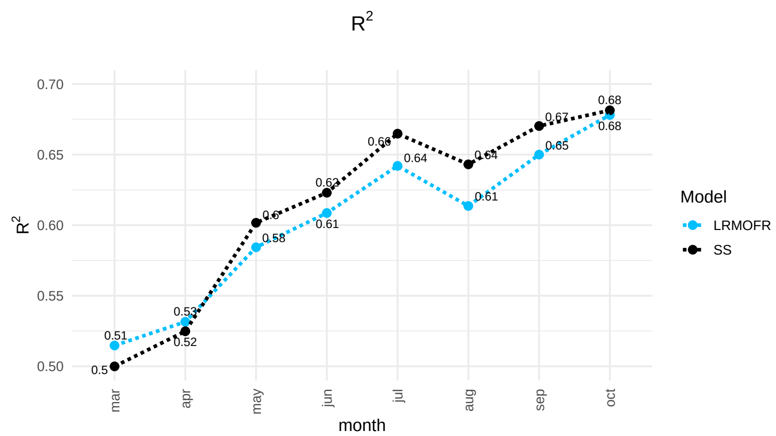
Figure 1 shows that the RMSE of the state space model is generally lower than that of the linear regression model, except for the first two and the last month.



08-02 v7.pdf

Figure 1. Root Mean Square Error (RMSE) of the two models with best RMSE. The graph shows the RMSE in the test classes for each month of the school year. LRMOFR: Linear Regression for each Month with an Optimal Forgetting Rate model. SS: State Space model.

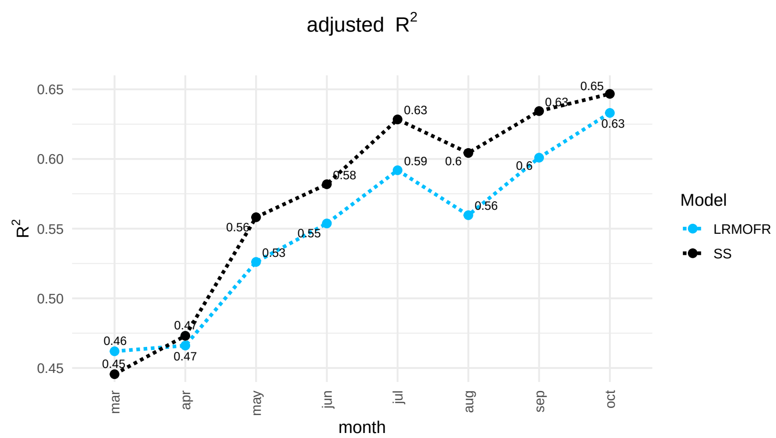
Figure 2 compares R^2 for the LRMOFR model with R^2 for the SS model. We see again that in most months the state-state model has better R^2.



08-02 v7.pdf

Figure 2. R^2 of the two best models for each month of the school year.

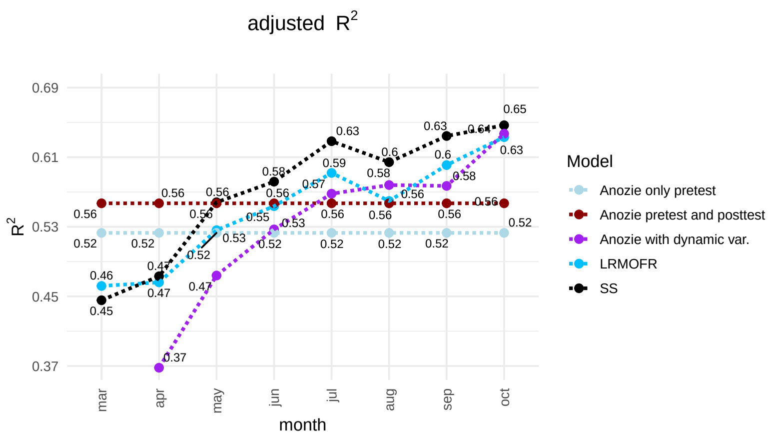
Since the state space model has much fewer parameters, adjusted R^2 is even better for the state space model (figure 3).



08-02 v7.pdf

Figure 3. Adjusted R^2 of the two best models for each month of the school year.

The use of adjusted R^2 allows comparison with models developed by other authors in other educational systems. For example, with the models of Anozie et al [1]. We have changed the month in order to be able to compare with those of the Chilean school year. As shown in figure 4, the adjusted R^2 of the State Space model is superior to the other models.



08-02 v7.pdf

Figure 4. Adjusted R^2 of the two best models for each month of the school year.

3.2. Optimal values of the parameters

For the regression model, we performed 200 cross-validations for each month. The results are in table 3. The parameters are quite stable.

Parameter	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct
ρ	0.71 (0.05)	0.71 (0.05)	0.71 (0.05)	0.71 (0.05)	0.71 (0.05)	0.71 (0.05)	0.71 (0.05)	0.71 (0.05)
γ_τ	33.16 (2.22)	23.78 (2.24)	18.77 (2.14)	17.17 (2.16)	14.56 (2.17)	17.38 (1.91)	15.87 (1.70)	15.34 (1.59)
β_0	240.04 (3.45)	240.04 (3.45)	240.04 (3.45)	240.04 (3.45)	240.04 (3.45)	240.04 (3.45)	240.04 (3.45)	240.04 (3.45)
$\beta_{1,\tau}$	-11.80 (1.75)	-11.49 (2.04)	-11.38 (1.94)	-9.24 (1.73)	-9.26 (1.72)	-10.27 (1.76)	-8.96 (1.71)	-8.62 (1.65)
$\beta_{2,\tau}$	5.00 (1.12)	4.83 (1.05)	4.60 (0.95)	3.98 (0.95)	4.23 (0.83)	4.01 (0.93)	3.61 (0.95)	3.00 (0.89)
$\delta_{1,\tau}$	-1.00 (2.03)	1.73 (3.29)	7.93 (2.97)	5.57 (2.99)	6.75 (2.68)	7.22 (3.24)	12.21 (2.80)	12.01 (2.49)
$\delta_{2,\tau}$	-0.38 (1.73)	3.58 (1.63)	5.23 (1.71)	3.98 (2.25)	10.57 (1.85)	7.08 (2.08)	9.30 (1.90)	6.70 (2.23)
$\delta_{3,\tau}$	-5.69 (1.63)	-2.30 (2.12)	0.52 (2.06)	5.88 (2.31)	7.18 (2.71)	7.03 (2.69)	6.86 (2.74)	4.61 (2.40)
$\delta_{4,\tau}$	-2.08 (1.89)	-0.66 (2.12)	4.14 (2.24)	8.00 (2.67)	4.55 (2.90)	2.61 (2.66)	4.45 (2.35)	3.18 (2.94)
$\delta_{5,\tau}$	1.83 (1.64)	5.19 (1.92)	2.28 (2.51)	7.76 (1.48)	13.82 (2.35)	5.16 (2.54)	8.07 (2.56)	11.09 (2.30)
$\eta_{1,\tau}$	4.70 (1.69)	14.64 (2.61)	13.34 (2.22)	15.76 (2.33)	14.85 (2.07)	9.85 (2.59)	5.28 (2.54)	3.03 (2.19)
$\eta_{2,\tau}$	-	1.94 (1.21)	2.57 (0.99)	0.71 (1.52)	-0.82 (1.87)	1.82 (2.90)	2.13 (2.56)	7.64 (2.52)
$\eta_{3,\tau}$	4.10 (1.90)	1.20 (2.74)	1.86 (1.67)	-3.29 (3.02)	-6.90 (3.66)	-4.86 (3.17)	-3.59 (3.13)	0.50 (3.52)
$\eta_{4,\tau}$	-	1.21 (1.58)	2.59 (1.75)	1.59 (2.05)	3.49 (2.03)	10.44 (2.89)	5.24 (2.10)	3.55 (2.02)
$\eta_{5,\tau}$	-	-2.62 (1.83)	1.26 (2.46)	-2.72 (1.55)	-6.25 (2.42)	-0.79 (2.08)	1.09 (3.20)	1.23 (2.54)

Table 3. Estimated parameters for the Linear Regression for each Month with Optimal Forgetting Rate model obtained with 200 cross-validations.

For the state space model, we also performed 200 cross-validations, but not separated by month. The results are in table 4. Again, the parameters are quite stable. We found that A is basically 1. That is, the state space is a pure accumulator.

Parameter	Mean	SD
x_0	258.3901	3.4421
A	1.0010	0.0006
K	0.0835	0.0085
$A-K$	0.9175	0.0085
$Q^{\frac{1}{2}}$	0.8982	0.1080
$R^{\frac{1}{2}}$	10.4133	0.5388
B_1	0.3787	0.2285
B_2	0.5888	0.1422
B_3	0.0770	0.1739
B_4	0.1984	0.1855
B_5	0.6834	0.1631
B_6	1.1153	0.2337
B_7	0.3170	0.1014
B_8	-0.0979	0.1365
B_9	0.4526	0.1363
B_{10}	-0.0675	0.1544
B_{11}	24.6766	2.0893
B_{12}	-14.0852	2.0651
B_{13}	6.143	1.2942

Table 4. Estimated parameters for the State-Space model obtained with 200 cross-validations

A is 1, but K is 0.08. Therefore, $A-K$ is 0,92. Thus the predictor uses a forget rate that is close to 8% from one month to the next one. The standard deviation $R^{\frac{1}{2}}$ of the random forcing term $w(t)$ is low and equal to 0.9 SIMCE points. The standard deviation $Q^{\frac{1}{2}}$ of the measurement noise $v(t)$ is about 10 SIMCE points, which is in the order of that reported by the SIMCE year-end national statistical test.

For the accuracy indicators, the highest parameter B that translate accuracies in SIMCE points is for the Data and Probabilities strand. However, it does not stand out much from the others. For the other performance indicator, the difference between the number of questions answered correctly on the first attempt and the rest of the questions, the Number and Operatives strand has a B above the rest of the strands. It is a much larger B than the B's of the rest of the strands. That is expected given the largest number of exercises done belongs to the Numbers and Operations strand, and also to the fact that the year-end national standardized test has a lot more questions from that strand. There are some values of B that are negative. This may be because it is perhaps convenient to do fewer exercises of those strands that have less weight in the end-of-year test, and then spend more time practicing exercises of the other strands.

3.3. Contribution of the different variables to the prediction

If we expand the state space expression, we can compare term by term with the LRMOFR model as sown in table 5.

Effect	LRMOFR	SS
Intercept	β_0	$A^{T-\tau-1}(A-K)^{\tau}\widehat{x}(0)+A^{T-\tau-1}\sum_{t=1}^{\tau}(A-K)^{\tau-t}Ky(t)$
Pretest	$\gamma_\tau p_i$	$B_{11}\left(A^{T-\tau-1}(A-K)^{\tau}u_{11}(0)\right)$
'My behavior is a problem...'	$\beta_{1,\tau}\Theta_{i,j}^{(1)}$	$B_{12}\left(A^{T-\tau-1}(A-K)^{\tau}u_{12}(0)\right)$
'Math is easy for me'	$\beta_{2,\tau}\Theta_{i,j}^{(2)}$	$B_{13}\left(A^{T-\tau-1}(A-K)^{\tau}u_{12}(0)\right)$
Accuracy	$\frac{\sum_{t=1}^{\tau}\sum_{k=1}^{5}\rho^{\tau-t}\delta_{k,\tau}a_{i,j,k,t}}{\sum_{t=1}^{\tau}\rho^{\tau-t}}$	$\sum_{k=1}^{5}B_k\left(A^{T-\tau-1}\sum_{t=1}^{\tau}(A-K)^{\tau-t}u_k(t) + \sum_{t=\tau+1}^{T-1}A^{T-t-1}\widehat{u}_k(\tau,t)\right)$
Difference	$\frac{\sum_{t=1}^{\tau}\sum_{k=1}^{5}\rho^{\tau-t}\eta_{k,\tau}d_{i,j,k,t}}{\sum_{t=1}^{\tau}\rho^{\tau-t}}$	$\sum_{k=6}^{10}B_k\left(A^{T-\tau-1}\sum_{t=1}^{\tau}(A-K)^{\tau-t}u_k(t) + \sum_{t=\tau+1}^{T-1}A^{T-t-1}\widehat{u}_k(\tau,t)\right)$

Table 5. Term-by-term comparison of each of the effects of the LRMOFR model with the SS model

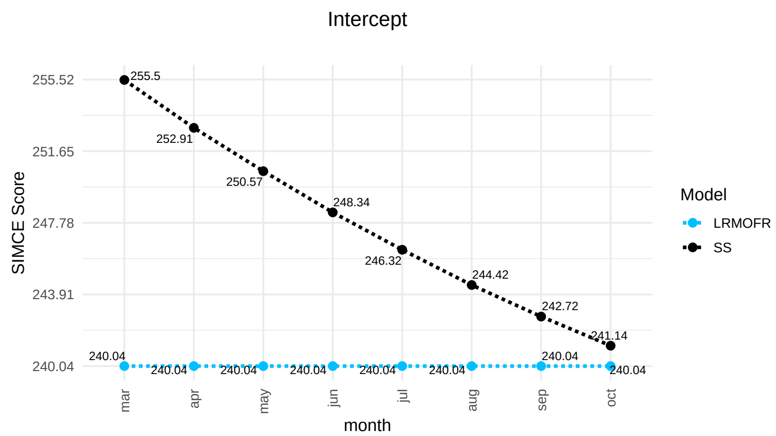
We first analyze the intercept. It is the prediction made each month if we did not have any data on the student. In the case of the LRMOFR model, we impose the same value for all months. In the SS model there is an initial estimate that falls but is corrected with the observations of each month. The expression for SS is

$$A^{T-\tau-1}(A-K)^{\tau}\widehat{x}(0)+A^{T-\tau-1}\sum_{t=1}^{\tau}(A-K)^{\tau-t}Ky(t)$$

Since A is 1, this expression is:

$$(A-K)^{\tau}\widehat{x}(0)+\sum_{t=1}^{\tau}(A-K)^{\tau-t}Ky(t)$$

We see in figure 5 that the values of the intercept of both models are similar and in the last months they coincide.



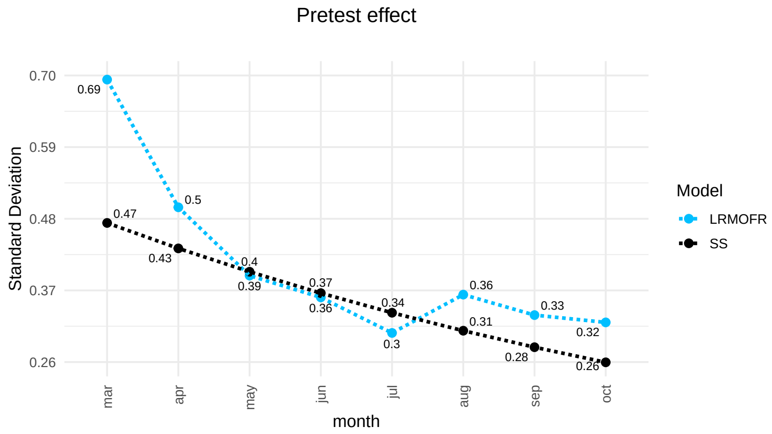
08-02 v7.pdf

Figure 5. Intercept on the LRMOFR and SS models for each month of the school year.

Next, we analyze the effect of the pretest. We increase the pretest on standard deviation and compute its effect on the prediction. In both models, in the first months the effect of the pretest is much higher than in the later month of the school year (figure 6). For the state space model, the effect is monotonically and exponentially decreasing. In march, one standard deviation increase in the pre-test causes an increase of 0.54 standard deviations in the prediction of the SIMCE score, whereas in October causes an increase of 0.28 standard deviations. This means that the effect decreases to one half. In the sequence of linear regressions, the effect is also decreasing but rapidly reaches a plateau in May. The graph shows the values of the parameter γ_τ presented in Table 3 and the values of the expression

$(A-K)^{\tau}B_{11}$ with the value of the parameters presented in Table 4 but scaled to standard deviations of the SIMCE test score.

The decay of the effect of the pretest is natural and expected. Given that the platform constantly accumulates the performance of hundreds of exercises performed by the student, then as the times passes by, the pretest information is decreasingly important to predict the result of the end-of-year test. Only at the beginning of the year the pretest is a highly valuable information. What these models add is the quantification of the decay. In the case state space model the decays is at a rate of 91.75% from one month to the next. This is a loss of more than 8% per month.

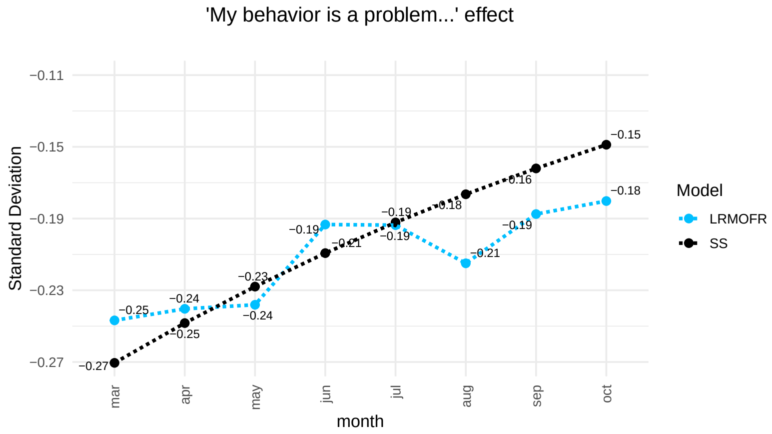


08-02 v7.pdf

Figure 6. Pretest effect. This is the contribution on the SIMCE prediction of one standard deviation increase of the pre-test.

At the classroom level, the greatest effect was one that is a proxy for the classroom climate. It is the average of the responses of all the students in the class to the degree of agreement in a scale from 0 to 5 with the statement that "My behavior is a problem for the teacher". At the beginning of the school year, in March, an increase in one standard deviation in this response generates a decrease of 0.27 standard deviations in the prediction of the SIMCE score of each student in the course (Figure 7). By the end of the year, this negative contribution decreases by approximately half. As Figure 7 shows, the decay of this classroom climate effect in the state-space model is exponential, while in the linear regression model it is more complex. Moreover, it basically reaches a plateau in June. In the state space, again the decay from one month to the next one of the classroom climate is of a 91.75%. This is the same as the decay of the pretest. This means that in each variable the contribution to the end-of-year test drops by 8% from one month to the next. Figure 7 shows the values of the parameter $\beta_{1,\tau}$ presented in Table 3 and the values of the expression $(A-K)^{\tau}B_{12}$ with the value if the parameters presented in Table 4, but scaled to standard deviations of the SIMCE test score.

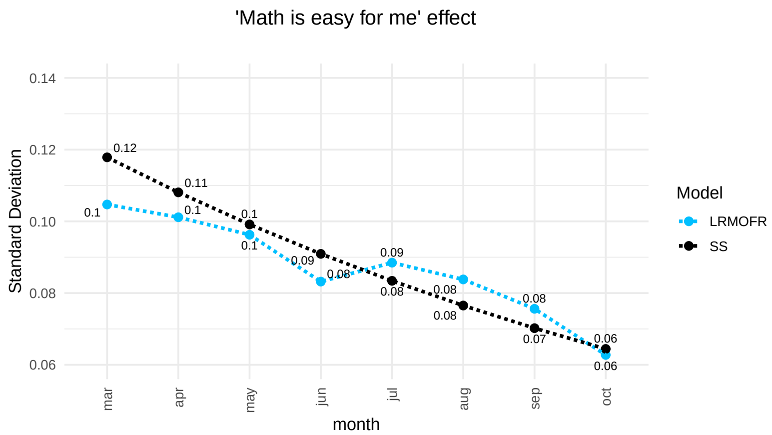
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2022 doi:10.20944/preprints202208.0124.v1



08-02 v7.pdf

Figure 7. Effect of the class average of the degree of agreement with "My behavior is a problem for the teacher". This is the contribution to the SIMCE prediction of an increase in one standard deviation of the student belief.

The personal variable that contributes the most to the prediction of the SIMCE score is the belief that mathematics is easy for me (Figure 8). This is the degree of agreement on a scale of 1 to 5 with "Mathematics is easy for me". At the beginning of the school year, in March, a student's one standard deviation increase in that belief causes a 0.12 standard deviation increase in the predicted SIMCE score. However, as the school year progresses, this contribution decreases. At the end of the year, the contribution of this belief reaches half of what it had in March. As Figure 8 shows, the decay of this effect in the state space model is exponential, while in the linear regression model it is more complex. In the state space, the decay of this belief is again a 91.75% from one month to the next, the same as the decay of the pretest and of the class climate. Figure 8 shows the values of the parameter $\beta_{2,\tau}$ presented in Table 3 and the values of the expression $(A-K)^{\tau}B_{13}$ with the value if the parameters presented in Table 4 but scaled to standard deviations of the SIMCE test score.



08-02 v7.pdf

Figure 8. Effect of the "Math is easy for me" subjective belief. This is the contribution to the SIMCE prediction of an increase in one standard deviation of the student belief.

We now compare the effects of deliberate practice of exercises from each of the 5 strands of the national curriculum. In each month we have two performance metrics for the five strands. On the one hand, we have the accuracy. This is the rate of exercises answered correctly on the first attempt. On the other hand, we have the difference between the number of exercises answered correctly on the first attempt and the number of exercises answered correctly on other attempts.

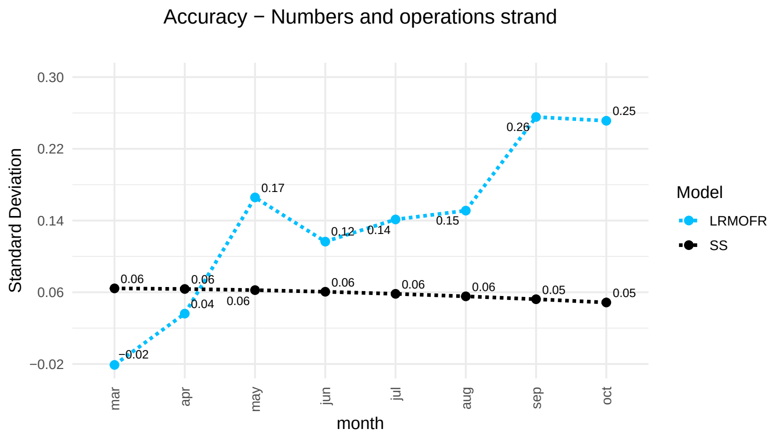
For the LRMOFR model, Table 3 shows the effects on the prediction of the end-of-year test of a one standard deviation increase in the accuracies. These effects are measured by the parameters δ for each curriculum strand and for each month. In general, the effects are positive. The first month of the school year is March. However, this is a month with very little activity since the implementation started in the fourth week of that month. In general, the Numbers and Operations strand is the one with the greatest effect. Students do most of the exercises corresponding to this strand. The percentage of exercises that the students did that belong to this strand is 58.78%. On the other hand, this strand is the one with the most items in the end-of-year SIMCE test. The second strand with the greatest effect is Patterns and Algebra, which has the greatest effect in July. Then there is the Measurement strand which is the one with the most effect in June. In July and August, the Geometry strand has a great effect. Only in October, just before the SIMCE test, does the Data and Probability strand have the higher effect. This pattern is typical in fourth grade classes in Chile. Data and Probability is taught at the end of the year.

On the other hand, an increase in one standard deviation in the difference between the number of questions answered correctly on the first attempt and the number of questions answered correctly on other attempts generates far more effect on the Strand of Numbers and Operations. This large effect is shown in table 3 in the η parameters. The only different month is the final month, October, where the Patterns and Algebra strand has the highest effect.

Reviewing Table 4 and comparing the effect values between the LRMOFR and SS models, we see that the effects of the LRMOFR model decay at the rate $\rho = 0.71$. This decay is much faster than the one in the SS model, whose decay rate is $(A - K) = 0.9175$. On the other hand, in the SS model, the effect of the difference in the number of exercises answered correctly on the first attempt and the rest of the exercises is much greater in the Numbers and Operations strand than in the rest of the strands. But also, that effect is much greater than that of accuracy. Since increasing accuracy can require much more effort, the large effect of this difference suggests that in many cases it may be worthwhile to increase the number of exercises. For example, if we have an accuracy of 0.7 with 10 exercises, then the difference is 7-3 = 4. If we go up to 20 exercises while maintaining the accuracy, then the difference is 14-6 = 8. Since B6 is more than three times B1, and since the variables are normalized, then the performance in terms of the difference between questions answered correctly in the first attempt and the number of those not answered correctly in the first attempt has 3 times the effect of the accuracy. That is, increasing one standard deviation in the difference has three times the effect of raising the accuracy by one standard deviation.

We show in figures 9 and 10 only the case of the Numbers strand, which is by far the Strand with the highest percentage of exercises performed. The percentage is 65% of the exercises performed in the year.

Figure 9 shows the effect of a student who is one standard deviation above the mean in accuracy each month. In the SS model we assume that in the future the student will perform one standard deviation above the mean.

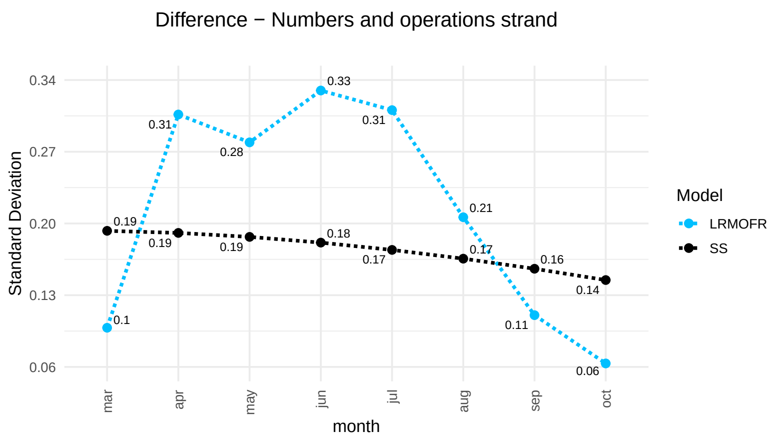


08-02 v7.pdf

Figure 9. Effect of an accuracy one standard deviation above the mean on every month. This is the contribution to the SIMCE prediction of an increase in one standard deviation of the student accuracy.

Figure 10 shows the effect of a student who in each month is one standard deviation above the mean in the difference between the number of correct answers in the first attempt and the number of correct answers in another attempt, but assuming that the accuracy is equal to the mean. This means that the student has done many more exercises, but with accuracy equal to the average of all the students. For example, if the average accuracy is 0.7, then if he has done 7 good exercises on the first attempt and 3 correct exercises on other attempts, it has an accuracy of 0.7 and a difference of 4. Whereas if he has done 70 good exercises on the first attempt and 30 correct ones on other attempts, then he also has an accuracy of 0.7 but the difference is 40.

In the SS model we assume that in the future the student will continue with a performance of one standard deviation above the mean.



08-02 v7.pdf

Figure 10. Effect of a difference one standard deviation above the mean on every month. This is the contribution to the SIMCE prediction of the difference between number of exercises solved correctly in the first attempt and the number solved correctly on other attempts.

3.4. The optimal control problem: ask students to do exercises more carefully or do more exercises

In each session the teacher must decide how many exercises to place and what level of performance will be acceptable. If students make a lot of mistakes, then she may need to stop the practice and explain the core concepts in more detail. She can also assign student monitors to explain and give help to peers. She can also ask the students to do the exercises more carefully and seek for help. There may be students responding very quickly and even randomly.

That is, the teacher must decide whether to give more importance to the percentage of correct answers in the first attempt or to the total number of exercises attempted by the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2022 doi:10.20944/preprints202208.0124.v1

students. For example, she has to decide between getting the class to perform 30 exercises per student with 70% of them solved correctly at the first attempt, or to perform only 20 exercises but with 80% them solved correctly at the first attempt. Her dilemma is: Which strategy causes better long-term learning? Or, more precisely, which strategy of exercise practicing causes a prediction of a highest score in the at the end-of-the-year national test?

Consider that we are in the month of τ. That is, with complete information on the students up to that month. And that for a following month s, the teacher is planning to carry out exercises of the strand k. Therefore, if a student achieves that month the performance given by $\widehat{u}_k(\tau,s)$ and $\widehat{u}_{k+5}(\tau,s)$, then the predictor of the SS model year-end test result will add the effect $\Delta_{k,\tau,s}$

$$\Delta_{k,\tau,s} = A^{T-s-1}(B_k\widehat{u}_k(\tau,s) + B_{k+5}\widehat{u}_{k+5}(\tau,s))$$

In terms of the accuracy a, the number of exercises solved correctly at the first attempt N, and number of exercises solved correctly in other attempts M, the effect is:

$$\Delta_\tau(N_{k,s}, M_{k,s}) = A^{T-s-1}\left[B_k\left(\frac{\frac{N_{k,s}}{N_{k,s}+M_{k,s}} - \mu^a_{k,\tau}}{\sigma^a_{k,\tau}}\right) + B_{k+5}\left(\frac{N_{k,s} - M_{k,s} - \mu^d_{k,\tau}}{\sigma^d_{k,\tau}}\right)\right]$$

with mean and deviation $\mu^a_{k,t}, \sigma^a_{k,t}$ for $a_{k,t}$, and $\mu^d_{k,t}, \sigma^d_{k,t}$ the mean and deviation for $d_{k,t}$, $\forall t \in \{1,\ldots,T-1\}$.

Rewriting in terms of accuracy and total number of exercises L, the effect on the predictor of an accuracy a, and a number of exercises L to be performed in a month s after the current month τ is given by

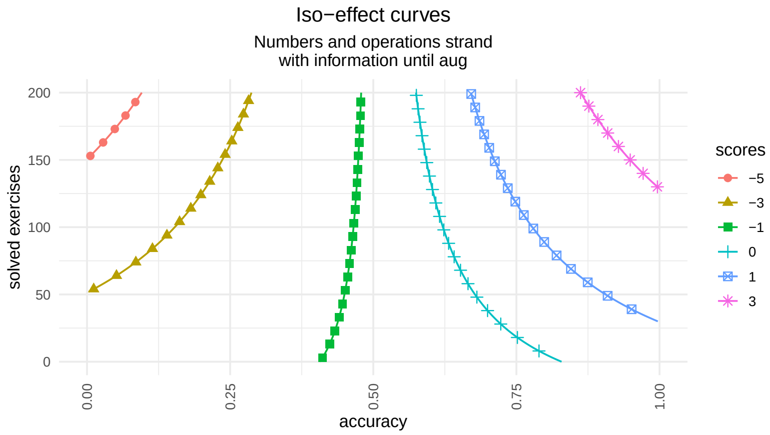
$$\Delta_\tau(a_{k,s}, L_{k,s}) = A^{T-s-1}\left[B_k\left(\frac{a_{k,s} - \mu^a_{k,\tau}}{\sigma^a_{k,\tau}}\right) + B_{k+5}\left(\frac{L_{k,s}(2a_{k,s}-1) - \mu^d_{k,\tau}}{\sigma^d_{k,\tau}}\right)\right]$$

Then for a level δ, the pairs of accuracies and number of exercises L for a strand k with the effect δ on month s posterior to τ, the last month with student data, is given by:

$$a_{k,s} = \frac{\delta A^{s-T+1} + \frac{B_k\mu^a_{k,\tau}}{\sigma^a_{k,\tau}} + \frac{B_{k+5}(L_{k,s}+\mu^d_{k,\tau})}{\sigma^d_{k,\tau}}}{\frac{B_k}{\sigma^a_{k,\tau}} + \frac{2L_{k,s}B_{k+5}}{\sigma^d_{k,\tau}}}$$

This expression describes a curve in the $a-L$ plane. Figure 11 shows the iso-effect curves as a function of the accuracy a, and the number of exercises L. Since A is basically 1, the iso-effect curves do not depend on the month s. However, they do depend on τ, through the means and standard deviations of the accuracy and the difference.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2022 doi:10.20944/preprints202208.0124.v1



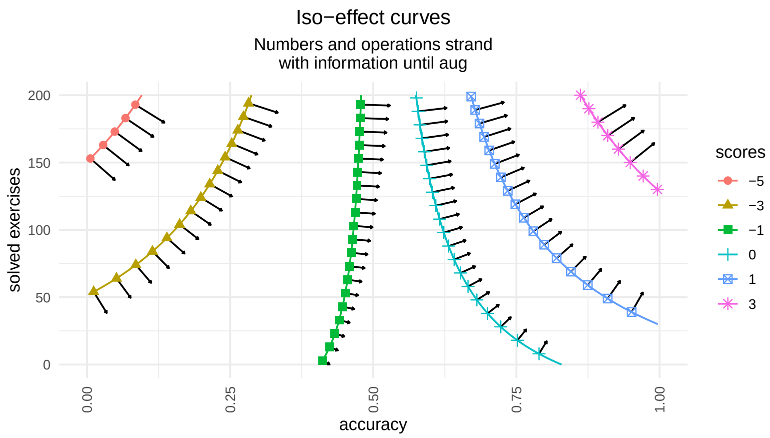
vector 08-01 v2.pdf

Figure 11. Iso-effect curves for the Number and Operations strand, in September, where the last month with student data is August.

The gradient of the effect with respect to a and L is:

$$\nabla\Delta_\tau(a_{k,s}, L_{k,s}) = \begin{pmatrix} \frac{\partial\Delta_\tau}{\partial a}(a_{k,s}, L_{k,s}) \\ \frac{\partial\Delta_\tau}{\partial L}(a_{k,s}, L_{k,s}) \end{pmatrix} = \begin{pmatrix} A^{T-s-1}\left[\frac{B_k}{\sigma^a_{k,\tau}} + \frac{2L_{k,s}B_{k+5}}{\sigma^d_{k,\tau}}\right] \\ A^{T-s-1}\left[\frac{B_{k+5}(2a_{k,s}-1)}{\sigma^d_{k,\tau}}\right] \end{pmatrix}$$

Figure 12 shows the vector field of the gradients superposed to the iso-effect curves for the Number and Operations strand for the month September when we have information up to August.



vector 08-01 v2.pdf

Figure 12. Effect gradients and iso-effect curves for the Number and Operations strand, in September, where the last month with student data is August.

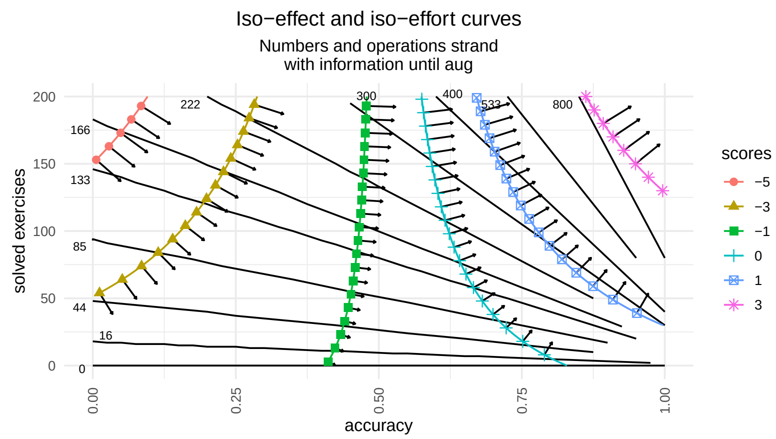
From figure 12 we see that if the accuracy is below 50%, then it is convenient to reduce the number of exercises to be done but to do them more carefully to improve the accuracy. For example, if we have been doing 130 exercises with an accuracy of 30%, then in order to increase the result of the year-end test, it is better to lower the number of exercises and improve the accuracy. This is due to the fact that in the coordinate (0.30, 130) of the plane $a - L$ the gradient points to have higher accuracy a, but to do less number of exercises L.

Let us consider that performing a number L of exercises with an accuracy a means an average effort $e(a, L)$ for the students of the class. It is reasonable, within a range, that by doubling the number of exercises to be performed in a session then the effort doubles. But the dependence of effort on accuracy is more complex. Let's assume the following dependency:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2022 doi:10.20944/preprints202208.0124.v1

$$e = \frac{cL}{(1.1 - a)}$$

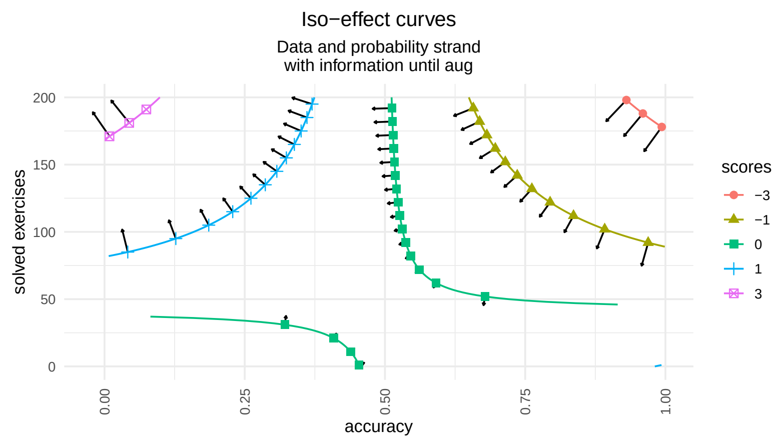
Then, for a given strand of the curriculum, the iso-effort curves are those of figure 13. By considering the iso-effect curves or the vector field of gradients, for each strand of the curriculum, for a given level of effort, we can find the optimal combination of accuracy and number of exercises to perform in the session.



vector 08-01 v2.pdf

Figure 13. Iso-effort, effect gradients, and iso-effect curves for the Number and Operations strand in September, where the last month with student data is August.

In figure 13 we can see that for an effort less than or equal to 166, in the Numbers and Operation strand, the best combination is to plan to do 40 exercises and with an accuracy of 95%. At that point on the $a - L$ plane, the effect gradient is perpendicular to the iso-effort curve.



vector 08-01 v2.pdf

Figure 14. Iso-effort curves and effect gradients for the Data and Probability strand in September, where the last month with student data is August.

Figure 14 shows that in the Data and Probability strand the gradient always points to lower accuracy. One possible explanation is that this may be because it is more effective to spend that effort on other strands.

4. Discussion

Improving the quality of education depends on the quality of transmission of pedagogical practices. Like any cultural phenomenon, improvement depends on the access to reliable performance indicators. Good indicators of teaching practices are those that achieve good long-term learning in students. It is therefore necessary for each teacher to be able to estimate and understand the effect of teaching practices on long-term learning. The teacher can then use these estimates to make teaching decisions for every lesson. Therefore, teachers need to be able to predict the effect of applying, in the remaining months, the different options of teaching strategies. In this paper we consider for each month and curriculum strand two options: how many exercises and with what level of accuracy.

To solve this problem, teachers face several challenges. First, given that they prefer to pose their own exercises, this generates a large number of questions, each one attempted by a small number of students. Thus, we couldn't use models based on big data such as deep learning.

Second, the questions of the national end-of-year tests are unknown to the teachers and there is no public information available on the results of students in each of those questions. For this reason, predicting the results on those tests is much more complex than in other tests. Strategies to predict students' performance in a question based on data of the performance of many similar students in similar questions, such as the algorithms used by Matrix or Tensor factorization or Deep Learning Networks, cannot be directly applied.

Third, another difficulty with predicting long-term learning is that it is not the same as short-term learning. It is not only a phenomenon of decaying and forgetting. It is much more complex. According to [27] during teaching, what we can observe and measure is performance, which is often an unreliable index of long-term learning. There are strategies that lead to less learning than others in the short term, but generate more learning than the rest over in the long term.

With the information from the responses of 500 students from 24 courses who completed an average of 1089 closed-question exercises on the ConectaIdeas platform during the year 2017, we built two types of models: linear regressions and state space models. We chose the best option of variables for each of them. We selected those variables analyzing information from 18 courses and we calculated the prediction errors in the remaining 6 courses. We carried out this cross-validation process 200 times. With these results we can now answer affirmatively the three research questions.

Research Question 1: To what extent can teacher designed questions (low or no stake quizzes) help predict students' long-term learning as measured by end-of-year standardized state tests?

Despite having questions designed by teachers and each of them answered by few students, the linear regression and state-space models achieve predictions with errors comparable to those of the literature [1,20]. We found that with both types of models, we were able to improve the month-to-month predictions. Additionally, the state space model is in general better than the linear regression model. This better performance is very important because it achieves it with much less parameters. The state-space model has only 17 parameters to fit, whereas the linear regression has 106 parameters. That is, the SS model has only 16% of the parameters of the linear regression model. This means that it must be more robust and generalizable to other schools and years.

Research Question 2: To what extent a hidden Markovian hidden state of the student accumulated knowledge up to the current month can be estimated so that along with the probable deliberate practice on the next months are sufficient enough to predict with good accuracy the end-of-year standardized state tests?

The SS model is a simple accumulator that adds monthly performance. From month to month, the predictor adds the score for the new month to the score for previous months. Additionally, the mechanism of the SS model adds the score that it estimates for the following months according to the strategy that the teacher plans to follow. In the first month, the mechanism includes a base score, and the contribution of three sources of information: a pretest, beliefs of each student, and estimation of the climate of the class. The predictor translates these three contributions into three scores, and adds them. Thus, the teacher has the initial prediction in view, and with the performance of each month, she has an additive combination to predict long-term learning.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2022 doi:10.20944/preprints202208.0124.v1

Both models estimate the impact of the pretest and how its impact decreases as the year goes by. This expected, since as the ConectaIdeas platform accumulates information on student performance then the initial information is less relevant. Likewise, they estimate the impact of students' beliefs about mathematics. This is very relevant for the teacher as it gives her a clue to determine the importance of this emotional factor and how much effort she should put into dealing with it. Both models also estimate the impact of the class climate on the prediction of each student's end-of-year test. Again, dimensioning this effect is very relevant for the teacher. It guides her to quantify how much effort she should dedicate to improving the climate of the class. In both cases, these effects also decline over the course of the year. In the case of the state-space model, the three effects decay at the same rate. Both models also estimate the effect of the practice of exercises in each of the five strands of the curriculum. The teacher can thus compare the effect of performance on each strand on the prediction of each student results on the end-of-year national test. This helps her decide how much intensity to devote to each strand of curriculum.

A central element of the models is the information for each month. It has two performances: the accuracy, which is the percentage of questions answered correctly on the first try. The other element is the difference between the number of questions answered correctly on the first attempt and the rest of the questions. They are very meaningful and very common elements in the daily practice of the teachers. The predictor mechanism just adds those contributions.

A particular characteristic of the SS model is that it includes the effect of strategies on future months. If we use the typical pattern of the student shown in previous months, the predictor uses that pattern and achieves a good prediction in the national end-of-year test. But the teacher can also try different options for the following month.

Research Question 3: To what extent a state-space model can help teachers to visualize the tradeoff between asking students to do exercises more carefully or doing more exercises, and thus help drive whole classes to achieve long-term learning targets?

The state-space model provides a very transparent and simple prediction mechanism. First, it has very few parameters. They are just what is needed. As there are five strands in the Chilean mathematics curriculum, and in each one we have two performance metrics, this already brings together 10 parameters. They are parameters that translate both performances in the five strands into the corresponding contribution to the test score. The rest are the parameters translate into the test score the contribution of the pretest, the beliefs about ease in mathematics, and the climate of the class. Finally, there is the rate of forgetting from one month to another.

Second, the whole mechanism is an additive one. The score of each contribution is added with the rest.

Third, the peculiarity of the state space model is that from one month to another, the mechanism updates the prediction in a Markovian way. It is enough to have the predictor of the last month, not the whole history of information. This is a huge advantage; as last month's predictor sums it all up. It is not necessary to go to gather information from previous months.

Fourth, for each month and for each strand, the mechanism is a tool that helps the teacher can see the effect of her decisions. She can visualize ahead the effect of choosing how many exercises to request and what minimal accuracy to request. In a 2D graph the teacher can see the iso-impact curves of these two decisions.

Fifth, the teacher can also visualize the arrows of the gradients. They are an intuitive visualization that indicate how to move to achieve more impact in the national test at the end-of-the-year. This allows the teacher to design her optimal strategy for driving the entire class toward the goal of greatest long-term learning. It is not a black box recommending what to do. It is a tool that allows the teacher to simulate and visualize the effect of different strategies.

Sixth, the teacher can also superimpose the iso-effort curves in order to find an optimal solution. That is, she can seek the combination of the number of exercises and accuracy in

 doi:10.20944/preprints202208.0124.v1

order to achieve the greatest long-term learning while maintaining a limited and predefined level of effort. That is, not only the state-space model gives the teacher an understanding of the factors involved and how they contribute to the result of the end-of-year summative test. It also gives her a tool to simulate the effect of various practice strategies. It is a very natural and an intuitive way to visualize the trade-off between asking students to do exercises more carefully or doing more exercises. In summary, in this paper we have built a state-space model with control variables that allow the teacher to visualize an optimal control strategy in order to conduct sessions that achieve the best long-term learning.

Although there are applications of optimal control to student learning strategies [13,18,21,28], these are problems formulated with abstract situations. They do not include empirical data on classes with students, models that fit those data, and end-of-year national standardized tests that independently measure long-term learning. To the best of our knowledge, this is the first time that the problem faced by each teacher is presented as an optimal control problem, with whole year empirical data of hundreds of students in several classrooms, and where a very practical solution is proposed. Moreover, we have developed a graphical tool, intuitive and easily understandable, that help teachers visualize the effect of strategies. With this state-space formulation and with the iso-impact and iso-effort curves, at each session the teacher can review what her students have achieved in previous sessions, and decide how to drive the entire class from now on.

There are several aspects to investigate and develop in the future. One is to include an interaction between strands of the curriculum. In a previous study, we empirically identified the effect of some topics on others [2]. We plan to include this interaction in the state-space model. Another aspect is to incorporate the effect of written answers to open questions [8,31]. We have already done it for a linear regression model. We plan to include this information in the state-space model. There is also the effect of different strategies that the teacher can choose for deliberate practice. For example, the effect of spacing and interleaving [26]. Another important topic is to include the effect of conducting peer reviews of the written response to open-ended questions. Finally, there is the effect of peer tutors [9]. We plan to include these strategies in future state-space models.

Author Contributions: Conceptualization, R.A.; methodology, R.A.; software, O.U.; validation, R.A., and O.U.; formal analysis, R.A. and O.U.; investigation, R.A. and O.U.; resources, R.A.; data curation, R.A. and O.U.; writing—original draft preparation, R.A.; writing—review and editing, R.A.; visualization, R.A. and O.U.; supervision, R.A.; project administration, R.A.; funding acquisition, R.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by ANID/PIA/Basal Funds for Centers of Excellence FB0003

Data Availability Statement: Not applicable

Acknowledgments: Support from ANID/PIA/Basal Funds for Centers of Excellence FB0003 is gratefully acknowledged. We also thank the International Development Research Center (IDRC) and the Interamerican Development Bank (IDB) for supporting the RCT project where the data was obtained, Julián Cristia from IDB, Abelino Jiménez and Raúl Gormaz from Universidad de Chile for many suggestions for analysis.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

LRMOFR	Linear Regression for the Month with an Optimal Forgetting Rate
SS	State Space
RMSE	Root Mean Squared Error
SD	Standard Deviation
SIMCE	Sistema de Medición de la Calidad de la Educación (Mesurement System of the Quality of Education)
GPA	Grade Point Average
OLS	Ordinary Least Squares

References

1. Anozie, N.; Junker, B. Predicting end-of-year accountability assessment scores from monthly student records in an online tutoring system. In Proceedings of the AAAI Workshop on Educational Data Mining. Boston, Massachusetts, USA, 16 & 17 of July 2006.
2. Araya, R. & Van der Molen, J. Causal Dependence among Contents Emerges from the Collective Online Learning of Students, ICCCI 2013, vol 8083, pp. 641-650, Springer. doi: 10.1007/978-3-642-40495-5_64
3. Araya, R. & Van der Molen, J. Impact of a blended ICT adoption model on Chilean vulnerable schools correlates with amount of on online practice. Proceedings of the Workshops at the 16th International Conference on Artificial Intelligence in Education AIED 2013. Memphis, USA, 9-13 July, 2013.
4. Araya, R.; Gormaz, R.; Bahamondez, M.; Aguirre, C.; Calfucura, P.; Jaure. P.; Laborda, C. ICT supported learning rises math achievement in low socio economic status schools. In Proceedings of the 10th European Conference on Technology Enhanced Learning. Toledo, Spain, 15-18 September 2015. Lecture Notes in Computer Science, Volume 9307, pp 383-388. Springer. doi: 10.1007/978-3-319-24258-3_28
5. Araya R. Integrating Classes from Different Schools Using Intelligent Teacher Support Systems. IHSI 2018, vol 722. pp 294-300. Springer, Cham. doi: 10.1007/978-3-319-73888-8_46
6. Araya R. (2019) Teacher Training, Mentoring or Performance Support Systems?. AHFE 2018. vol 785, pp 306-315. Springer, Cham. doi: 10.1007/978-3-319-93882-0_30
7. Araya, R.; Arias Ortiz, E.; Bottan, N.; Cristia, J. Does Gamification in Education Work? Experimental Evidence from Chile. IDB WORKING PAPER SERIES Nº IDB-WP-982 Inter-American Development Bank, 2019. doi: 10.18235/0001777
8. Araya, R.; Diaz, K. Implementing Government Elementary Math Exercises Online: Positive Effects Found in RCT under Social Turmoil in Chile. Educ. Sci. 2020, 10, 244. doi: 10.3390/educsci10090244
9. Araya, R.; Gormaz, R. Revealed Preferences of Fourth Graders When Requesting Face-to-Face Help While Doing Math Exercises Online. Educ. Sci. 2021, 11, 429. doi: 10.3390/educsci11080429
10. Bagaloplan, A.; Zhang, H.; Hamidieh, K.; Hartvigsen, T.; Rudzicz, F.; Ghassemi, M. The Road to Explainability is Paved with Bias: Measuring the Fairness of Explanations. In proceedeings of the FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency. Seoul, Republic of Korea, 21-24 June 2022. doi: 10.1145/3531146.3533179
11. Bjork, R.A., & Bjork, E.L. Desirable difficulties in theory and practice. Journal of Applied research in Memory and Cognition, 2020, 9 (4), 475-479. doi: 10.1016/j.jarmac.2020.09.003
12. Bjork, E.; Boork, R. Making Things Hard on Yourself, But in a Good Way: Creating Desirable Difficulties to Enhance Learning. FABBS Foundation, Psychology and the real world: Essays illustrating fundamental contributions to society, 2016, pp. 56–64. Worth Publishers.
13. Chen Y.H. A revisit to the student learning problem. Optimal Control Applications and Methods, 2016, vol 12(4), pp 263–272. doi: 10.1002/oca.4660120405
14. Gervet, T.; Koedinger, K.; Schneider, J.; Mitchell, T. When is Deep Learning the Best Approach to Knowledge Tracing? Journal of Educational Data Mining, 2020, vol 12, No 3. doi: 10.5281/zenodo.4143614
15. Gezer, T.; Wang, C.; Polly, A.; Martind, C.; Pugaleee, D.; Lambert, R. The Relationship between Formative Assessment and Summative Assessment in Primary Grade Mathematics Classrooms. International Electronic Journal of Elementary Education, June 2021, vol 13, Issue 5, pp 673-685.
16. Heckman JJ. Skill formation and the economics of investing in disadvantaged children, Science, 2006, vol 312 (5782), pp 1900–1902. doi: 10.1126/science.1128898
17. Heckman, J.; Zhou, J. Interactions as Investments: The Microdynamics and Measurement of Early Childhood Learning, 24 october 2021
18. Lewis, D. Modeling student engagement using optimal control theory. Journal of Geometric Mechanics, 2022, vol 14 (1), pp 131-150. doi: 10.3934/jgm.2021032
19. Namoun, A.; Alshanqiti, A. Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Sys-tematic Literature Review. Appl. Sci. 2021, 11, 237. doi: 10.3390/app11010237
20. Pardos, Z.; Baker, R.; San Pedro, M.; Gowda, S.; Gowda, S. (2014). Affective States and State Tests: Investigating How Affect and Engagement during the School Year Predict End-of-Year Learning Outcomes. Journal of Learning Analytics, vol 1(1), pp 107–128. doi: 10.18608/jla.2014.11.6
21. Raggett, G.F.; Hempson, P.W.; Jukes, K.A. A Student-Related Optimal Control Problem, Bulletin of the Institute of Mathematics and Its Applications, 1981, vol 17, pp 133–136.

22. Raviv, L.; Lupyan, G.; Green, S.C. How variability shapes learning and generalization, Trends in Cognitive Sciences, 1 June 2022, vol 26(6), pp 462-483. doi: 10.1016/j.tics.2022.03.007
23. Rea, D; Burton, T. Does an empirical Heckman curve exist?, 2018. https://www.wgtn.ac.nz/__data/assets/pdf_file/0005/1716953/WP18-03-does-an-empirical-Heckman-curve-exist.pdf
24. Ruipérez-Valiente, J.A.; Muñoz-Merino, P.J.; Delgado, C. Improving the prediction of learning outcomes in educational platforms including higher level interaction indicators. Expert Systems, 17 July 2018. doi: 10.1111/exsy.12298
25. Rohrer, D.; Dedrick, R.; Hartwig, M. The Scarcity of Interleaved Practice in Mathematics Textbooks, Educ Psychol Rev 32, 10 January 2020, 873–883. doi: 10.1007/s10648-020-09516-2
26. Rohrer, D.; Dedrick, R.; Stershic, S. Interleaved Practice Improves Mathematics Learning. Journal of Educational Psychology, 2015. vol 107 (3), pp 900-908. doi: 10.1037/edu0000001
27. Soderstrom, N.; Bjork, R. Learning Versus Performance: An Integrative Review. Perspectives on Psychological Science, March 2015, vol 10(2), pp 176–199. doi: 10.1177/1745691615569000
28. Teklu S.W.; Terefe, B.B. (2022) Mathematical modeling analysis on the dynamics of university students animosity towards mathematics with optimal control theory. Scientific Reports. 12, 11578. doi: 10.1038/s41598-022-15376-3
29. Van der Molen, J. Minería de datos educacionales: Modelos de predicción del desempeño escolar en alumnos de enseñanza básica. En: Universidad de Chile, 2013.
30. Ulloa, O. Estimación de desempeño en evaluación sumativa, con base en evaluaciones formativas usando modelos espacio estado. MS Thesis. University of Chile, 2021.
31. Urrutia, F.; Araya, R. Do written responses to open-ended questions on fourth-grade online formative assessments in mathematics help predict scores on end-of-year standardized tests?, 2022 (submitted).
32. Zheng, G.; Fancsali, S.; Ritter, S.; Berman, S. Using Instruction-Embedded Formative Assessment to Predict State Summative Test Scores and Achievement Levels in Mathematics. Journal of Learning Analytics, 2019, vol 6(2), pp 153 —174.
33. Zytek, A.; Arnaldo, I.; Liu, D.; Berti-Equille, K.; Veeramachaneni, K. The Need for Interpretable Features: Motivation and Taxonomy. ACM SIGKDD Explorations Newsletter, June 2022, vol 24 (1), pp 1–13. doi: 10.1145/3544903.3544905