Using Embedded Formative Assessment to Predict State Summative Test Scores

Stephen E. Fancsali
Carnegie Learning, Inc.
501 Grant St., Ste. 1075
Pittsburgh, PA, 15219, USA
sfancsali
@carnegielearning.com

Guoguo Zheng
Yanyan Tan
University of Georgia
Athens, GA, USA
{ggzheng, yanyan.tan25}
@uga.edu

Steven Ritter
Susan R. Berman
Carnegie Learning, Inc.
501 Grant St., Ste. 1075
Pittsburgh, PA, 15219, USA
{sritter, sberman}
@carnegielearning.com

April Galyardt[#]
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA, USA
agalyardt@gmail.com

ABSTRACT

If we wish to embed assessment for accountability within instruction, we need to better understand the relative contribution of different types of learner data to statistical models that predict scores on assessments used for accountability purposes. The present work scales up and extends predictive models of math test scores from existing literature and specifies six categories of models that incorporate information about student prior knowledge, socio-demographics, and performance within the MATHia intelligent tutoring system. Linear regression and random forest models are learned within each category and generalized over a sample of 23,000+ learners in Grades 6, 7, and 8 over three academic years in Miami-Dade County Public Schools. After briefly exploring hierarchical models of this data, we discuss a variety of technical and practical applications, limitations, and open questions related to this work, especially concerning to the potential use of instructional platforms like MATHia as a replacement for time-consuming standardized tests.

CCS CONCEPTS

• **Applied computing**~Education~Computer-assisted instruction

LAK'18, March 7–9, 2018, Sydney, NSW, Australia

ACM 978-1-4503-6400-3/18/03…$15.00
https://doi.org/10.1145/3170358.3170392

KEYWORDS

Intelligent tutoring systems, formative assessment, mathematics education, accountability, assessment, predictive modeling

ACM Reference Format:

S.E. Fancsali, G. Zheng, Y. Tan, S. Ritter, S.R. Berman, and A. Galyardt. 2018. Using Embedded Formative Assessment to Predict State Summative Test Scores. In *LAK'18: International Conference on Learning Analytics and Knowledge, March 7–9, 2018, Sydney, NSW, Australia. ACM, New York, NY, USA*, 10 pages. https://doi.org/10.1145/3170358.3170392

[#] April Galyardt was a member of the faculty at the University of Georgia during the period of time in which she contributed to the presented work.

1 INTRODUCTION

Formative assessment of student learning is vital to guiding effective instruction (cf. [35]), and summative assessments are essential supports for accountability systems throughout the United States and the world. However, assessment of learning via traditional, high-stakes tests has come under fire for being misaligned with instructional goals and for taking time away from instruction. In addition to being predicated on outdated assumptions about learning [33], summative assessments crowd out valuable instructional time. Large school districts recently surveyed by The Council of Great City Schools [18] reported that, over a typical academic year, the average eighth grader spent 25.3 hours taking 10.3 tests, only considering district-administered tests. Assuming 180 instructional days per year and approximately six hours' instructional time per day, this already consumes over 2% of instructional time [18], which is further constrained by various others facets of day-to-day school activities (e.g., school assemblies, student discipline), making that percentage an exceedingly conservative estimate. Preparation for these exams consumes additional instruction time.

Public backlash against testing reflects skepticism about the alignment of testing and instruction and frustration with the

amount of instructional time devoted to testing [27]. In New York state, 27% of students opted out of high-stakes math testing in 2017 [22]. In Minneapolis, so many students have refused to take 10th and 11th grade state math exams that the state no longer recommends relying on exam results to estimate student performance, growth or achievement gaps among students [34]. There is a clear demand for new approaches to testing.

Recognition of the problems with standardized testing is not new. Nearly twenty years ago, Grigorenko and Sternberg [17] had already provided an extensive review of existing literature on "dynamic testing," contrasting dynamic tests with "static tests," and placing dynamic testing within the more general framework of dynamic assessment. Such approaches embed assessment within the learning process:

"[I]nstead of quantifying the existing set of of abilities and level of knowledge and viewing them as a basis for predicting children's subsequent cognitive development, dynamic testing has as its aim the quantification of the learning potential of the child during the acquisition of new cognitive operations" [17].

Further, they point out historical antecedents for approaches that assess students while they learn to at least the early 20th century. For example, they point to Binet [6] as both the creator of what they call static testing and as an advocate of process assessment. Grigorenko and Sternberg summarize Buckingham's 1921 view [10] to be "that the best measure of intelligence is one that takes into account the rate at which learning takes place, the products of learning, or both" [17].

Various recent approaches take seriously what Campione and Brown [11] call metrics for "dynamic testing", like the extent to which students ask for hints and how long it takes students to answer questions, using them as components of statistical models to predict outcomes on various standardized tests. Adaptive learning systems already collect such data as part of their formative assessment function. If we can do a reasonable job of predicting standardized test scores from these indicators of student learning, such scoring might serve as a replacement for standardized tests.

As a first step, demonstrating the validity and reliability of such predictive models is necessary to achieving the goal of replacing high-stakes standardized testing with innovative solutions that provide assessments as students learn. We begin by briefly describing Carnegie Learning's MATHia instructional platform. Then we review recent approaches to predicting standardized test scores with data from the ASSISTments system as well as Carnegie Learning's Cognitive Tutor technology [30], on which the MATHia platform is based. Next, we explain the dataset for the present study and our model specification approach. Finally, we provide our results, discuss these results as well as their limitations, and present avenues for future research in predictive modeling to support innovative assessment.

Our results and discussion focus on the relative contribution of factors that are readily available to a system that is expected to provide on-going formative assessment of student learning versus those factors which are not always available (or appropriate) to such a system but that may be available for retrospective analysis (e.g., prior year test scores, socio-demographic information, etc.) Importantly, the latter category also includes elements of the inherent hierarchy in data like these (e.g., class and school identity), which we consider using appropriate models.

2 MATHia™ + Cognitive Tutor™

MATHia is an intelligent tutoring system that is part of Carnegie Learning's middle school and high school blended mathematics curricula. Based on Cognitive Tutor [30] technology, MATHia is a fundamental, instructional component of the blended mathematics curriculum (for Algebra I) that was the subject of one of the most rigorous effectiveness studies ever done with such a mathematics curriculum [24]. Carnegie Learning's blended model calls for a 60%-40% split between student-centered, non-computer-based instructional time and time with the MATHia instructional platform, respectively.

Students learn in MATHia by solving multi-step, real world problems, which engage a variety of problem solving modalities (e.g., equation solving, proofs, graphing, word problems, etc.), organized into topical "workspaces." MATHia is based on the idea of mastery learning [7]; in each workspace, there are multiple, fine-grained knowledge components (KCs; or skills) [21] that students master before moving on to the next workspace. Students have multiple opportunities to learn each KC within a workspace, and KC mastery is tracked using Bayesian Knowledge Tracing (BKT) [12].

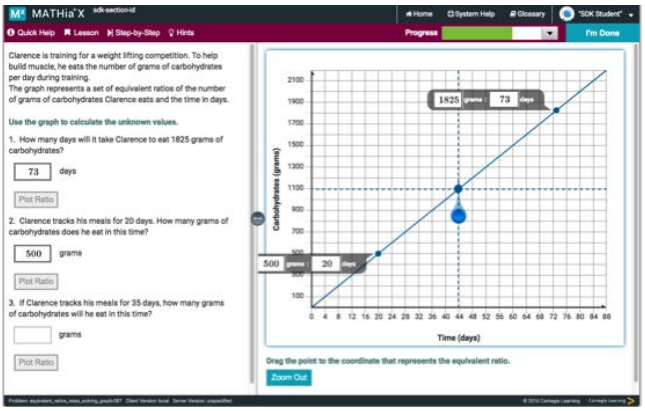

Figure 1: Screenshot from Carnegie Learning's MATHia intelligent tutoring system, based on Cognitive Tutor technology. (© Carnegie Learning, Inc.)

3 PRIOR WORK

Before considering the approach of the present project using MATHia data, we consider similar efforts using data from the

ASSISTments system to predict mathematics standardized test scores in Massachusetts. Similar efforts at to create predictive models for innovative assessment in domains outside of mathematics have been pursued in domains like reading (e.g., [5]) and physics (e.g., [32]), but even a brief review of such work is beyond the scope of this paper.

3.1 Using ASSISTments Data

Numerous papers [1, 2, 14, 20, 26] have considered various elements of process data (e.g., % correct on various types of items, metrics like hint-seeking, number of KCs mastered) from the ASSISTments system [28] to incorporate in regression and other predictive models of the Massachusetts Comprehensive Assessment System (MCAS) exam for math. More recent work with ASSISTments data has predicted standardized state test scores [25] relying on the predictions of so-called "detector" models of behavior and affective states [4].

Work by Feng, et al. [14] had access to item-level data for the MCAS and reported predictive accuracy in terms of the mean absolute difference (MAD) of predicted values for raw MCAS scores compared to a learners actual raw scores. Using "online testing metrics" similar to the process variables we consider and a stepwise linear regression approach similar to one of our approaches, [14] report a within-sample MAD of 5.533 points over a sample of 600 ASSISTments learners in 2004-2005. This MAD represents 10.25% error given the 54-point total possible raw MCAS score. In the present study we do not have access to raw, item-level data. However, Pardos, et al. [25] report both MAD and correlations of their predicted values on the MCAS to learners' actual MCAS scores using models that incorporate elements of learners' behavior (e.g., gaming the system [3]), affective states [4] (e.g., boredom), and performance (e.g., correctness on ASSISTments items). Reported correlations for models over three training and testing regimes range from 0.694 to 0.765.

3.2 Using Cognitive Tutor Data

Table 1 provides six sets of variable categories we will consider in this work. This model-labeling schema follows that provided by Ritter, et al. [31], with one additional category. In general, this previous work considered similar sets of variables within a single academic year as predictors of the US State of Virginia's SOL exam over a sample of 2,018 students in Grades 6-8. The Northwestern Education Alliance's (NWEA) Measures of Academic Progress (MAP) computer adaptive test was used as a pre-test in all analyses. Process variables (*process*) in [31] are a subset of variables we consider in this work. Socio-demographic variables (*demog*) are broadly similar to those considered in this work as well.

Table 2 provides for a comparison of the proportion of variance in SOL scores accounted for by linear regression models for Grade 7 learners, which rely on different sets of predictive variables. As variables were added from M1 (and M2) through M5, the Bayesian Information Criterion (BIC) [29] decreased, indicating a justification for increasing the complexity of the predictive models because of increases in predictability achieved.

Ritter, et al. [31] generalized models M1-M5 learned on Grade 7 data by testing these models on data from learners in Grade 6 and Grade 8. Perhaps unsurprisingly, model M5, which includes the most information about learners, achieved the greatest predictability. The authors achieved adjusted R^2 values similar to those in Table 2 for Grade 6 data. In fact, Model M5 tested on Grade 6 achieved a greater R^2 value ($R^2 = 0.62$) than on the training set. Nevertheless, the authors saw substantial decreases in predictability in Grade 8, likely due to the fact that the Grade 8 math population has a substantially different makeup than Grades 6-7 as students are tracked into Algebra I classes, leaving relatively weaker students taking Grade 8 math rather than Algebra I (data from which were not considered by [31]).

Table 1: Variable sets considered in this work (*pre-test* = pre-test score; *process* = process variables from MATHia usage; *demog* = demographic variables); M1-M5 are also considered by Ritter, et al. [31]; all sets include learner grade-level (6-8)

Model	Variable Sets
M1	*pre-test*
M2	*process*
M3	*process + demog*
M4	*pre-test + demog*
M5	*pre-test + demog + process*
M6	*pre-test + process*

Table 2: Summary of model fits for predicting Virginia SOL in Grade 7 based on different sets of variables (cf. Table 4 in [31])

Model	# Vars	BIC	R^2
M1	1	2041.5	0.5
M2	5	2181	0.43
M3	7	2167.8	0.45
M4	3	2030.6	0.51
M5	8	1928.4	0.57

Later work due to Joshi, et al. [19] adapted this model to data from a school district in West Virginia in a single academic year to predict math scores on the WESTEST 2 standardized test. While similar sets of variables were significant in these predictive models, such models did not achieve the level of predictability achieved by previous efforts [25, 31] (R^2 of the best model ~0.32).

Overall, models reported in Table 2 for the Virginia SOL compare favorably with models reported by Pardos, et al. [25] for MCAS. Recall that correlations between predicted and actual MCAS scores in that work ranged from 0.694 to 0.765, which correspond to approximate R^2 values from 0.482 to 0.585.

While such efforts, including predictive models that explain up to 62% of variability in a (held-out test set of) standardized test scores, are a good start, in what follows, we find models that substantially improve upon these prior examples, both in terms of the scope of data considered and predictability achieved.

4 DATA

The data we consider at present allow us to substantially scale up previous analyses, evaluating models by testing them on larger samples, in a new state (with a different standardized test), on data across academic years (and two different standardized tests used within the state over the time period of interest). We consider data from learners in Grades 6-8 in Miami-Dade County Public Schools in the US state of Florida who used MATHia over the course of four academic years. Sample sizes are reported in Table 3. Miami-Dade County Public Schools is one of the largest school districts in the United States.

Table 3: Sample sizes by grade-level (Gr) and academic year (13-14 = 2013-2014, etc.)

Gr	13-14	14-15	15-16	Total
6	2,914	2,471	3,542	8,927
7	3,827	3,596	3,505	10,928
8	1,200	1,301	1,018	3,519
All	7,941	7,368	8,065	23,374

In each year, the school district provided grade-level (i.e., 6-8), current year standardized test scores, previous year standardized test scores (*pre-test*), and socio-demographic data that could then be mapped to usage data from the MATHia system. In 2013-14, the mathematics component of the Florida Comprehensive Assessment Test (FCAT) constitutes the standardized test scale score, while in subsequent years the state adopted the Florida Standards Assessment (FSA) as their exam, so that exam constitutes the measure of interest. As in previous work, *pre-test* and end-of-year FCAT and FSA scores are standardized as z-scores for modeling, but we report statistical accuracy in terms of more interpretable FCAT and FSA scale score units.

The FCAT exam provides a developmental scale score ranging from 140 to 298 from Grades 3 to 8 [15], and the FSA exam provides a developmental scale score ranging from 240 to 393 from Grades 3 to 8 [16]. Both exams define five achievement levels. Levels 3-5 constitute "passing" the exam(s). Ranges of scale scores that are mapped to each achievement level vary from year to year and from grade-to-grade and are generally large for Levels 1 and 5, but the size of the range of scores for Levels 2, 3, and 4, which are important for determining whether a student passes or fails the exam, is nearly always between 11 and 15 points across years and grade-levels. This provides a crude, but useful, benchmark for thinking about the statistical accuracy of models we develop in the next section.

Socio-demographic variables (*demog*) considered (and most frequently occurring values) include:

- Ethnic Category (White, Hispanic, Black, & Other)
- Limited English Proficiency (LEP) Status (Enrolled, Not Enrolled, & Former)
- Exceptional Student Education (ESE) Status (Gifted, Learning Disability, & Other)
- Free/Reduced-Price Lunch (FRPL) Status (a rough socio-economic status indicator: Free, Reduced, & Denied)

The most frequently occurring of these socio-demographic categories are transformed into binary dummy variables (representing membership in the category) that are included in models.

We consider process variables drawn from the set of variables considered by previous work with Cognitive Tutor predicting Virginia's SOL exam [31] as well as several novel variables. As in previous work (cf. Figure 2 in [31]), process variable distributions often had a long right tail, justifying the use of a log-transformation to make distributions (approximately) normal. Further, since there is often wide variability in content and/or features of a workspace and corresponding variability in student work within these workspaces, we follow this previous work by standardizing several variables: transforming the variable, for each workspace, into a z-score which represents the difference (in units of standard deviation) between a particular student's value of a process variable within a workspace and the mean value of that variable across all students who worked in that workspace. For these variables that are "standardized within each workspace," we calculate a single value for a student by taking the mean of z-scores across the workspaces in which students worked. Variables that are aggregated over the entire academic year are standardized as such (i.e., a z-score is calculated for each variable according to a student's work with respect to the global mean for all students across the entire year). For those variables that were both log-transformed and standardized within each workspace, the log-transformation preceded standardization (and calculation of a mean z-score across workspaces per student).

Process variables (*process*) considered (and transformations applied to them) include:

- Workspaces mastered per hour: number of workspaces from which learners graduate (by mastering all KCs) per hour (log-transformed)

- Problems per workspace (log-transformed and standardized within each workspace)
- Number of KCs mastered (log-transformed)
- Total problem solving time (log-transformed)
- Assistance per problem (i.e., hints requested + errors committed per problem) (log-transformed and standardized within each workspace)
- Workspaces encountered (log-transformed)

Notably, the amount of learner usage in the 2013-2014 academic year was lower than in the two subsequent years. Median learner problem solving in 2013-2014 was approximately 20.7 hours while in 2014-2015 and 2015-2016 median usage was approximately 31.6 hours and 30.3 hours, respectively.

5 MODEL SPECIFICATIONS & LEARNING

As noted, we consider pre-test scores (*pre-test*) and categories of process variables (*process*) and socio-demographic variables (*demog*), progressively, as we specify and learn models, seeking to better understand the relative contributions of these categories of variables to (and the significance of individual variables in) successful predictive models of standardized test scores.

In what follows, we describe two methods of specifying non-hierarchical models (using stepwise linear regression and random forest methods) to predict standardized test scores and then consider using additional data about schools in which learners were enrolled to consider hierarchical models of this data. After we discuss the predictive success of different sets of variables within these models, we compare and contrast the predictive results and practical utility of models that account for the inherent hierarchy of this type of data (e.g., students working within schools) versus the models in this section that do not explicitly[1] do so.

5.1 Non-Hierarchical Models

Despite the previous success of linear regression approaches to this problem (e.g., [14, 31]), we compare a stepwise (ordinary least squares) linear regression approach to a random forest approach [9], which may be appropriate in the case that linearity and parametric assumptions of normality fail in the present data.

5.1.1 Stepwise Linear Regression (SLR). We use a stepwise procedure to find the best ordinary least squares regression model for six candidate sets of predictors (i.e., to find models M1 through M6). For each candidate set of predictors, the maximal set of model variables includes all the possible two-way interaction terms and the quadratic terms of process variables if there are any. Starting from a model that includes all variables in the set (but no interaction or quadratic terms), a single variable from the current model is removed or a single variable from the maximal set of model variables is added at each step that will decrease the BIC most, repeating this procedure until no single variable can be added or removed to further decrease the BIC. The choice of BIC provides for simpler models in terms of the number of variables that will be included, hopefully providing for better generalizability for held out data sets.

5.1.2 Random Forest (RF). Random Forest (RF) models [9] are learned by an ensemble machine learning method that partners decision tree learning with bootstrap aggregation (i.e., "bagging") [8] to produce models that are less likely to over-fit training data and consequently achieve better accuracy on held-out test data. Such models often perform well even in cases in which relationships among variables are non-linear and parametric assumptions like normality may be violated. As such, random forests provide an appropriate foil to our relatively simple linear regression methods, and we compare predictive accuracy achieved by each method.

Since the outcome variable of interest in our case is continuous, the method we deploy is called random forest regression. Learning such an ensemble model proceeds by inferring a large set of decision trees using bootstrap samples with replacement of subsets of training data. Predictions made by each individual decision tree are averaged, allowing each learned model to contribute to the overall predicted outcome. Individual decision tree learning proceeds by making recursive binary cuts on the predictors and dividing the predictor space into a set of hyper-rectangles. Observations that fall within the same hyper-rectangle will be assigned the same predicted value of the outcome variable, which is the average of all the cases in that hyper-rectangle. The cutting rule is to minimize the total sum of squared residuals across hyper-rectangles. The learning process stops when the hyper-rectangles include less than 5 cases.

5.1.3 Reporting Statistical Accuracy & Predictability. We consider the accuracy of the best models found by the procedures above for the variable sets corresponding to M1 through M6 on training data from the 2013-14 school year in terms of their mean absolute deviation (MAD) and root mean square error (RMSE) (see Table 4) in predicting learners' FCAT and FSA scores in terms of raw scale score points:

$$MAD = \frac{1}{n}\sum_{i=0}^{n} |FSA_i - \widehat{FSA_i}| \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=0}^{n} (FSA_i - \widehat{FSA_i})^2}, \quad (2)$$

[1] One might argue that the inclusion of various socio-demographic variables in some of the models does provide for some limited, implicit representation of this hierarchy (e.g., that within a large school district there are often clusters of students with a particular socio-economic status, etc.).

where $\widehat{FSA_i}$ is a particular model's predicted FSA score for learner *i*, and FSA_i is the actual FSA score for learner *i* (substituting FCAT for FSA in (1) and (2) when testing on 2013-14 data). In addition, for its heuristic utility as a measure of variability accounted for by a linear model as well as to compare our results to previous work reported in §3, we report models' R^2 values.

5.1.4 Questions for Analysis. We present our results around providing answers to the following three questions:

- *Q1*: Are SLR models or RF models more accurate in predicting test scores?
- *Q2*: What accuracy can be achieved?
- *Q3*: How does accuracy vary over models 1-6?

We focus discussion on models trained on data from the 2013-14 school year. However, this discussion also applies to the same analysis using 2014-15 data for training and data from 2013-14 and 2015-16 for testing; results suggest no significant deviations in the pattern of results found using 2013-14 as training data. This includes approximate accuracy achieved, differences in accuracy over models M1 through M6, differences between MAD and RMSE on the training and test data sets, etc. That is, answers to questions *Q1-Q3* remain roughly the same regardless of training set chosen. This is perhaps surprising given that 2014-15 data reflected more usage for the typical student in general than 2013-14 data, and the FCAT was used in 2013-14 while the FSA was used in subsequent years.

5.1.5 Results. Table 3 provides RMSE and MAD measures of accuracy on the training set. On both measures, SLR outperforms RF. We see the same pattern of SLR outperforming RF on both accuracy metrics when models were applied to held-out test data in Years 2014-15 and 2015-2016 (comparisons on test data omitted for brevity). We witness the same pattern when using 2014-2015 data for training (table omitted for brevity), so *Q1* is adjudicated in favor of SLR. Such results also support the assumption that the predictors and the outcome variable in the current study are linearly dependent.

Table 4: Comparison of the accuracy of the best models learned by stepwise linear regression (SLR) and random forest (RF) using RMSE and MAD on training data from 2013-14 school year

	RMSE		MAD	
Model	**SLR**	**RF**	**SLR**	**RF**
1	11.777	12.512	8.841	9.284
2	12.483	12.605	9.366	9.477
3	11.916	12.13	8.987	9.149
4	11.513	11.627	8.653	8.721
5	10.41	10.514	7.838	7.926
6	10.54	10.632	7.944	8.011

With respect to the accuracy our models achieve (*Q2*), we see that there are some differences in magnitude between RMSE and MAD, indicating some variance in the magnitudes of the errors in our predictions, but we see that both MAD and RMSE are below or within the range we indicated for different achievement levels (generally 10-15 points per level) in the FCAT and FSA, which indicates a great deal of promise that our models, at worst, can be expected to provide predictions of a student's achievement level within one level of that predicted by the FCAT or FSA. We provide scatterplots of predicted versus actual test score values in Figure 2 for the best model M5 on the training set (2013-14) as well as on the two test sets (2014-15 & 2015-16). Test data R^2 values, especially for M5 and M6, are also considerably larger than those reported in prior literature.

To begin addressing *Q3*, we find that relying on pre-test data alone (M1) provides for predictions that are approximately one point better than relying on process data alone (M2); the same holds true for training on 2014-15 data and testing on 2013-14 and 2015-16 data (cf. Table 5). Each of these "minimal" models can account for variability in held-out test data at a level comparable to the best models reported by Ritter, et al. on training data [31] and to the best models for ASSISTments tested on held-out data reported by Pardos, et al. [25].

Table 5: SLR results for model learned on 2013-14 data applied to test data from 2014-15 (14-15) & 2015-16 (15-16), expressed as RMSE and R^2.

	RMSE		R^2	
M	**14-15**	**15-16**	**14-15**	**15-16**
1	12.780	13.070	0.6035	0.647
2	13.775	14.188	0.5393	0.584
3	13.376	13.602	0.5656	0.6177
4	12.362	12.745	0.629	0.6643
5	11.319	11.327	0.689	0.7349
6	11.343	11.518	0.6326	0.7258

Adding demographic data to each of these provides for modest improvements in accuracy, but the best models we find include *pre-test*, *process*, and socio-demographic (*demog*) variables (M5). For a high-level view of our predictions with model M5, Figure 2 provides scatterplots of predicted values of FCAT and FSA scores against learners' actual or "true" values.[2] However, it is important to realize that the drop in accuracy from M5 to M6 (when we drop demographic data from consideration) is only 0.024 points in terms of RMSE on 14-15

[2] A reviewer points out that Figure 2 indicates that the residuals of our regression models are not strictly homoscedastic, especially due to individuals that have a true FCAT or FSA score that is either the maximum or minimum value. How to deal with these individuals, especially those with the minimum value, presents an interesting question for future research. Perhaps it is possible to develop statistical models that will point out specific characteristics of students that are likely to perform especially poorly (or possibly exceptionally well). If so, such models might be useful for targeting intensive intervention.

test data and 0.191 points in 15-16 test data. To the contrary, Table 5 shows that when using 2014-15 data for training, accuracy increases from M5 to M6 by 0.107 points when testing on 2013-14 data while accuracy decreases by 0.144 points when testing on 2015-16 data. Recall that accuracy measures are on the raw developmental score scale (with differences between achievement levels being from 10-15 points), so these are small differences indeed.

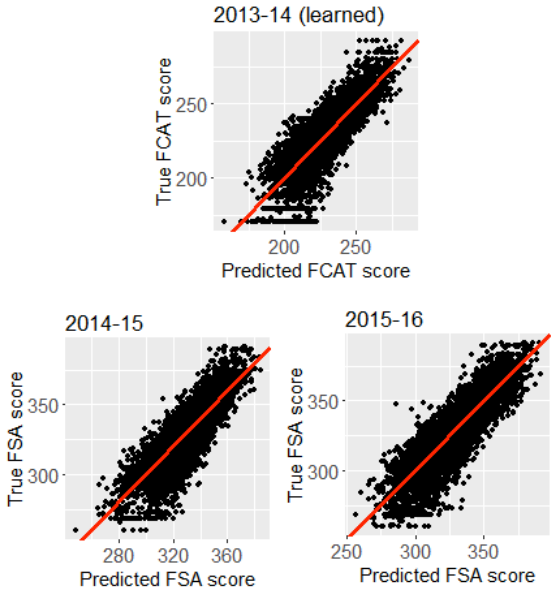

Figure 2: Scatterplots of predicted values of FCAT and FSA scores against actual scores from M5 learned using SLR on 2013-14 data and applied to test data from 2014-15 and 2015-16. The red line is a reference line from the graph origin with slope = 1. Table 5 presents details of this model.

Table 6: SLR results for model learned on 2014-15 data applied to test data from 2013-14 (13-14) & 2015-16 (15-16), expressed as RMSE and R^2.

	RMSE		R^2	
M	**13-14**	**15-16**	**13-14**	**15-16**
1	11.951	12.962	0.6001	0.6528
2	12.996	14.035	0.5271	0.593
3	12.758	13.587	0.5443	0.6185
4	11.864	12.668	0.6059	0.6684
5	10.952	11.444	0.6642	0.7294
6	10.845	11.588	0.6707	0.7225

Not relying on most socio-demographic data for predicting test scores is important from the standpoint of practical implementations of these predictive models. It is unacceptable for a system to handicap (or boost) a student's predicted score simply because of their status with respect to certain socio-demographic categories (e.g., ethnicity). In addition to concerns of explicit bias that could be introduced by relying on such socio-demographic categories, concerns of other forms of algorithmic bias (e.g., [13, 23]) must also be carefully considered and, if present, remediated to implement fair systems of assessment in the real world.

Nevertheless, for categories like exceptional student education, there are often various affordances or alternative tests made available that may make a different predictive model for some classes of students reasonable. Nevertheless, ideally we would rely on models like M2 that only include student process and performance data for the current year.

Insights from the *demog*, *process*, and *pre-test* variables found to be significant in M5 are worth considering. To provide a deeper dive into the models we consider here, we provide a summary of the regression model M5, including parameters estimates and significance, learned by SLR in Table 6.

Consistent with prior work explored in §3, *pre-test* and many *process* variables are significant in M5. Process variables that represent overall student progress (e.g., *Number of KCs mastered*) and efficient completion of material (e.g., *Problems per workspace*) are generally in the spirit of variables found by prior work on the Cognitive Tutor and ASSISTments platforms. Notably absent from this final model are *Workspaces mastered per hour* and *Assistance per problem*, both of which are prominent in models reported by Ritter, et al. [31].

Also consistent with existing literature, we find that a variety of socio-demographic variables are significant predictors of exam score. While variables that encode ethnicity are notably absent, free/reduced-priced lunch status variables are significant, yet are not necessarily of practical help in delivering fair, online predictions to educational stakeholders (e.g., a use case in which a predictive model in embedded within a product for progress monitoring).

However, the significance of ESE and LEP status indicators provide an important pointer to areas for future focus. Knowing that such status is a significant predictor of the exam score, it is possible that students with limited English proficiency and exceptional students (of various sorts) demonstrate proficiency and learning progress within MATHia in ways we have yet to fully understand. In the future we can seek out better performance indicators for such students to provide fairer, improved predictive models without relying explicitly on such a status indicator (i.e., to bolster M2 models, which really represent the overall goal of developing these predictive models). Construction of separate models for categories of exceptional students may also be feasible. That all said, we emphasize that dropping such socio-demographic categories (from M5 to M6) does not lead us to substantially worse predictive models of test scores.

Table 7: Estimated SLR model M5 learned on 2013-14 data; standardized parameter estimates; results of this model applied to data from 2014-15 and 2015-16 school years reported in Table 4; (*p < 0.001; **p < 0.01; *p < 0.05)**

Variable	Coefficient
Intercept	-2.254***
Pre-test	0.486***
Grade 8	0.557***
Grade 7	0.345***
Problems per workspace	0.311***
Pre-test x Problems per workspace	0.039**
Number of KCs mastered	0.192***
Total problem solving time	0.299***
Workspaces encountered	0.204***
Problems per workspace x Number of KCs mastered	-0.134***
(Total problem solving time)^2	-0.022***
LEP: Enrolled	-0.083***
LEP: Former	0.039**
ESE: Gifted	0.607***
ESE: Gifted x Workspaces Encountered	-0.112***
ESE: Learning Disability	-0.247*
ESE: Learning Disability x Workspaces encountered	0.065
ESE: Other	-0.433*
ESE: Other x Workspaces encountered	0.125*
FRPL: Free	-0.111***
FRPL: Reduced	-0.076**
FRPL: Denied	-0.053
LEP: Not Enrolled	0.130

5.2 Hierarchical Models

Modeling to this point has not explicitly accounted for the inherent hierarchy in educational data of this sort. Students learn within schools (and with teachers in classes), so we consider explicitly modeling this and the extent to which such considerations may improve predictability of test scores. We only consider models that include school identity, as school-level effects could at least plausibly carry over from year to year within a district. For example, school culture, classroom expectations, and student background may not drastically vary from year to year, but classes (and often teachers) change with sufficient frequency (i.e., in the case of classes, nearly always) that testing a model learned on one year's data with data from subsequent years would be difficult, involving some form of imputation or setting of appropriate prior distributions for parameters that would represent as-yet-unseen teachers or classes. Even in the present data set, schools present in the 2014-15 data set were not present in data from 2013-14, so we use data from 2014-15 to infer linear mixed effects models that can account for this school hierarchy. We then test these models on data from 2013-14 and 2015-16 and compare results to those obtained from SLR to see if explicitly modeling this hierarchy provides for significant boosts in accuracy.

5.2.1 Linear Mixed Effects Models. Linear mixed effects models take their name from the fact that they include both fixed and random effects, assuming that, for random effects, observed subpopulations have different values for coefficients in a linear model. In contrast, typical linear models like ordinary least squares regression models do not take subpopulations (or the hierarchical nature of the data) into account, and we only estimate a single set of parameter values for so-called fixed effects. Suppose, as is generally reasonable in educational data like this, that each school represents a different subpopulation. A random intercept and random slope model with predictors X_1, X_2, ..., X_s can be written as,

$$\begin{aligned} Y_{ij} = \beta_{0j} + \beta_{1j}X_{1i} + \beta_2 X_{2i} + \cdots + \beta_s X_{si} + \varepsilon_i \\ \beta_{0j} = \gamma_0 + \mu_{0j} \\ \beta_{1j} = \gamma_1 + \mu_{1j}, \end{aligned} \quad (3)$$

where Y_{ij} is the outcome variable of student i in school j, β_{0j} denotes the intercept of school j and β_{1j} denotes the slope of school j for predictor X_1. The rest of the predictors, X_2, ..., X_s have fixed effects only. This model has distribution assumptions, β_{0j} ~ N(γ_0, τ_{00}), β_{1j} ~ N(γ_1, τ_{11}) and ε_i ~ N(0, σ^2). Since we do not intend to provide an exhaustive consideration of hierarchical models, we do not estimate the covariance of β_{0j} and β_{1j}here.

In the current study, we fit a multilevel linear model for each candidate set of predictors, M1-M6, that has a random intercept per school as well as a random slope of previous FSA or FCAT score (if applicable) to model differential effects of learner prior knowledge across schools. Each multilevel model is built upon the corresponding best SLR model to provide the strongest comparison possible. That is, the only difference between the multilevel linear models and best SLR models are the random effects.

5.2.2 Results. Table 7 provides a comparison of model performance between the best SLR model trained on 2014-15 data and a linear mixed effects model that included, as fixed effects, all of the variables in the best SLR model as well as a random intercept per school and a random slope for *pre-test* per school. We see that, in terms of RMSE accuracy, the SLR models M5 and M6 outperform the hierarchical (i.e., linear mixed effects) model over both test data sets, indicating that some combination of *pre-test* and *process* variables appear to be promising as sufficiently informative to obviate the need to consider hierarchy explicitly. This is a boon for the practical application of these models, as it is likely not acceptable that school identity be explicitly considered in making predictions about learning for accountability. As we have also already noted, in large school

districts there can also be variability in terms of (the extent to) which schools use a particular platform from year-to-year, introducing possible difficulties if one is to rely on the identities of schools remaining stable across years for modeling.

Nevertheless, the linear mixed effects models do outperform SLR for M1-M4 in 2015-16 and in M2-M3 in 2013-14, indicating that models that use *process* variables alone, for example (e.g., M2), might benefit from this hierarchical approach (or investigations into how to capture these school-level effects with as-yet-unconsidered process variables).

Table 8: Comparison of the accuracy of the best models learned by stepwise linear regression (SLR) (cf. Table 6) and linear mixed effects modeling (LME) trained on data from 2014-2015 and tested on data from 2013-2014 and 2015-2016 using RMSE

	2013-14		2015-16	
Model	**SLR**	**LME**	**SLR**	**LME**
1	11.951	12.099	12.962	12.725
2	12.996	12.741	14.035	13.685
3	12.758	12.704	13.587	13.349
4	11.864	11.99	12.668	12.563
5	10.952	10.981	11.444	11.553
6	10.845	10.881	11.588	11.598

6 DISCUSSION & FUTURE WORK

Our results show the potential for using formative assessment as a replacement for end-of-year standardized tests. While the ability to predict test outcomes from the types of data we consider is likely necessary to eventually replace high-stakes exams, this ability remains insufficient to do so. Nevertheless, when such formative assessment is embedded within high-quality, effective instruction, the potential to increase instructional time and enhance learning outcomes is substantial. Our best predictions come from the most complete model: M5, using SLR on 2013-14 training data, achieves RMSE of 11.327 on 2015-16 test data, out-performing both SLR using 2014-15 training data and the linear mixed effects model built on 2014-15 training data. However, we see excellent results from the non-hierarchical M6, which does not take demographics into account (e.g., providing the best accuracy on the 2013-14 data as a test set when trained on 2014-15 data) and reasonable results from M2, a model that knows nothing about the student, other than performance within MATHia (achieving accuracy comparable within roughly one point to that of models based on pre-test/prior knowledge data alone).

Many statistical questions remain. Can an omnibus model learned over multiple years of data increase predictive accuracy on future data? Rather than include grade-level in the model, should we build separate models for each grade-level? Is it possible to use the same model across tests for multiple states (e.g., by calculating an internal score and translating this score into an appropriate scale score for each state)? How early in the academic year is it possible to reliably predict test scores from student work? If item-level data were available (or possibly more advanced statistics were reported), we could begin to establish (or know) upper bounds on predictability we could expect in the best case by considering split-half and other forms of reliability of the underlying standardized test (cf. [14]). Further, we have treated the prediction problem in the present work as a regression task to predict a continuous test score. How do the results and relative magnitude of statistical accuracies translate into a classification task? At least two possible classification tasks are apparent: predict whether a student is at or above level 3 (i.e., proficient) and predict the specific achievement level (1-5) into which a student falls. Would various machine learning classifiers applied to the data reveal difference variables (and categories thereof) as more or less important than what we report here?

Practical questions remain as well. To use such predictive models in real products deployed at scale like MATHia, how best can we represent an evolving prediction of student test scores based on their work as they make progress throughout the year? What is an easily interpretable way to represent uncertainty in those predictions? For instance, the model reported in Table 6 featuring a variety of performance variables and interaction terms provides for reasonable statistical accuracy, but it is not obviously and easily explainable to the end user of a system who may rely on its predictions or other school district stakeholders. How do we resolve tensions between reporting predictions about standardized test scores with assigning students grades based on their work within the instructional platform?

The benefits of using embedded formative assessment over end-of-year testing may be enormous. We estimate that, in the student population studied here, 40 classroom days (out of a 180-day school year) are currently devoted to standardized assessment (30 days for in-class benchmark exams and 10 days for final-test preparation and administration). Recovering that classroom time for instruction could have great impact. In addition embedded formative assessment serves a real-time instructional purpose of informing both students and teachers of progress towards end-of-year goals. Finally, a system using embedded formative assessment better supports personalized learning, in which students are assessed when they are ready, not on an arbitrary end-of-year date.

Solutions to these open statistical and practical questions have the potential to drive genuine innovation in assessment for accountability that can lead to increased instructional time, better personalized learning, and overall improved learning outcomes. We look forward to continuing investigations into these solutions.

ACKNOWLEDGMENTS

This work was supported by an Ignite Challenge Middle and High School Math Award to Carnegie Learning from the NewSchools Venture Fund. Carnegie Learning also thanks Miami-Dade County Public Schools and Ilia Molina for providing substantial amounts of data over the past 3+ years.

REFERENCES

[1] N.O. Anozie and B.W. Junker 2006. Predicting end-of-year accountability assessment scores from monthly student records in an online tutoring system. *American Association for Artificial Intelligence Workshop on Educational Data Mining (AAAI-06)*, July 17, 2006, Boston, MA.

[2] E. Ayers and B.W. Junker. 2008. IRT Modeling of Tutor Performance to Predict End-of-Year Exam Scores. *Educational and Psychological Measurement*, 68(6), 972-987.

[3] R.S.J.d. Baker, A.T. Corbett, I. Roll, K.R. Koedinger. 2008. Developing a generalizable detector of when students game the system. *User Model. User-Adap*, 18, 287-314.

[4] R.S.J.d. Baker, S.M. Gowda, M. Wixon, J. Kalka, A.Z. Wagner, A. Salvi,, V. Aleven, G.W. Kusbit, J, Ocumpaugh, L. Rossi. 2012. Towards sensor-free affect detection in Cognitive Tutor Algebra. In *Proceedings of the 5th International Conference on Educational Data Mining*, K. Yacef, O. Zaiane, A. Hershkovitz, M. Yudelson, J. Stamper (Eds.). 126-133.

[5] J.E. Beck, P. Lia, and J, Mostow. 2004. Automatically assessing oral reading fluency in a tutor that listens. *Technology, Instruction, Cognition and Learning*, 2(1-2), 61-81.

[6] A. Binet. 1909. *Les idées modernes sur les enfants*. [Modern concepts concerning children.] Flammarion: Paris.

[7] B.S. Bloom. 1968. Learning for Mastery. *Evaluation comment*, 1(2). UCLA Center for the Study of Evaluation of Instructional Programs, Los Angeles.

[8] L. Breiman. 1996. Bagging predictors. *Machine Learning*, 24(2), 123-140.

[9] L. Breiman. 2001. Random forests. *Machine Learning*, 45(1), 5-32.

[10] B.R. Buckingham. 1921. Intelligence and its measurement: A symposium. *Journal of Educational Psychology*, 12, 271-275.

[11] J.C. Campione and A.L. Brown. 1985. *Dynamic Assessment: One Approach and Some Initial Data*. Technical Report No. 361.Champaign, IL: University of Illinois at Urbana-Champaign, Center for the Student of Reading.

[12] A.T. Corbett, J.R. Anderson. 1995. Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Model. User-Adap*., 4, 253-278.

[13] D. Danks and A.J. London. 2017. Algorithmic bias in autonomous systems. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, C. Sierra (Ed.). 4691-4697. International Joint Conferences on Artificial Intelligence.

[14] M. Feng, N.T. Heffernan, and K.R. Koedinger. 2006. Predicting state test scores better with intelligent tutoring systems: developing metrocs to measure assistance required. In *Proceedings of the Eighth International Conference on Intelligent Tutoring Systems*, M. Ikeda, K. Ashley, and T-W. Chan (Eds.). Springer-Verlag: Berlin, 31-40.

[15] Florida Department of Education. 2014. *FCAT 2.0 and Florida EOC Assessments Achievement Levels*. Florida Department of Education: Tallahassee, FL. Retrieved 29 September 2017. http://www.fldoe.org/core/fileparse.php/3/urlt/achlevel.pdf

[16] Florida Department of Education. 2017. *Florida Standards Assessment: 2016–17 FSA English Language Arts and Mathematics Fact Sheet*. Retrieved 29 September 2017. http://www.fldoe.org/core/fileparse.php/5663/urlt/ELA-MathFSAFS1617.pdf

[17] E.L. Grigorenko and R.J. Sternberg. 1998. Dynamic Testing. *Psychol. Bull.* 124, 1 (1998), 75-111.

[18] R. Hart, M. Casserly, R. Uzzell, M. Palacios, A. Corcoran, and A. Spurgeon. (2015). *Student Testing in America's Great City Schools: An Inventory and Preliminary Analysis*. Council of Great City Schools, Washington, DC.

[19] A. Joshi, S.E. Fancsali, S. Ritter, T. Nixon. 2014. Generalizing and Extending a Predictive Model for Standardized Test Scores Based On Cognitive Tutor Interactions. In *Proceedings of the 7th International Conference on Educational Data Mining*, J. Stamper, Z. Pardos, M. Mavrikis, B.M. McLaren (Eds.). 369-370.

[20] B.W. Junker. 2006. Using on-line tutoring records to predict end-of-year exam scores: experience with the ASSISTments project and MCAS 8th grade mathematics. In *Assessing and modeling cognitive development in school: intellectual growth and standard settings*, R.W. Lissitz (Ed.). Maple Grove, MN: JAM.

[21] K.R. Koedinger, A.T. Corbett, C. Perfetti. 2012. The Knowledge-Learning-Instruction (KLI) framework: Bridging the science-practice chasm to enhance robust student learning. Cognitive Science, 36 (5), 757-798.

[22] S. Moses. 2017. State testing starts today; Opt Out CNY leader says changes are 'smoke and mirrors.' *Syracuse.com*. Retrieved 29 September 2017. http://www.syracuse.com/schools/index.ssf/2017/03/opt-out_movement_ny_teacher_union_supports_parents_right_to_refuse_state_tests.html

[23] C. O'Neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York: Crown Publishers.

[24] J.F. Pane, B.A. Griffin, D.F. McCaffrey, R. Karam. 2014. Effectiveness of Cognitive Tutor Algebra I at scale. *Educational Evaluation and Policy Analysis*, 36(2), 127-144.

[25] Z.A. Pardos, R.S.J.d. Baker, M.O.C.Z. San Pedro, S,M. Gowda, S.M. Gowda. 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*, 1(1), 107–128.

[26] Z.A. Pardos, N.T. Heffernan, B. Anderson, C. Heffernan. 2010. Using Fine Grained Skill Models to Fit Student Performance with Bayesian Networks. In *Handbook of Educational Data Mining*, C. Romero, S. Ventura, S. R. Viola, M. Pechenizkiy and R.S.J. Baker (Eds.). CRC Press, Boca Raton, FL. 417-426.

[27] PDK/Gallup. 2015. 47th annual PDK/Gallup Poll of the Public's Attitudes Toward the Public Schools: Testing Doesn't Measure Up For Americans. *Phi Delta Kappan*, 97(1).

[28] L. Razzaq, et al. 2005. The Assistment Project: Blending Assessment and Assisting. In *Proceedings of the 12th International Conference on Artificial Intelligence In Education*, C-K. Looi, G.I. McCalla, B. Bredeweg, J. Breuker (Eds.). 555-562. Amsterdam: IOS.

[29] G. Schwarz. 1978. Estimating the dimension of a model. *Ann. Statist.* 6(2), 461-464.

[30] S. Ritter, J.R. Anderson, K.R. Koedinger, A.T. Corbett. 2007. Cognitive Tutor: applied research in mathematics education. *Psychon. B. Rev*., 14, 249-255.

[31] S. Ritter, A. Joshi, S.E. Fancsali, T. Nixon. 2013. Predicting Standardized Test Scores from Cognitive Tutor Interactions.

[32] V.J. Shute and G.R. Moore. 2017. Consistency and Validity in Game-Based Stealth Assessment. In *Technology Enhanced Innovative Assessment: Development, Modeling, and Scoring From an Interdisciplinary Perspective*, H. Jiao and R.W. Lissitz (Eds.). Information Age Publishing, Charlotte, NC, 31–51.

[33] R.E. Snow and D.F. Lohman. 1989. Implications of cognitive psychology for educational measurement. In *Educational Measurement*, R.L. Linn (Ed.). 3rd Edition. 263-331. New York: American Council on Education/Macmillan.

[34] State of Minnesota, Office of the Legislative Auditor. 2017. *Standardized student testing: 2017 evaluation report*. Retrieved 29 September 2017. http://www.auditor.leg.state.mn.us/ped/pedrep/studenttesting.pdf

[35] D. Wiliam. 2011. *Embedded Formative Assessment*. Bloomington, IN: Solution Tree Press.