JOURNAL OF LEARNING ANALYTICS

SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH

Volume 6(2), 153 —174. http://dx.doi.org/10.18608/jla.2019.62.11

Using Instruction-Embedded Formative Assessment to Predict State Summative Test Scores and Achievement Levels in Mathematics

Guoguo Zheng[1], Stephen E. Fancsali[2], Steven Ritter[3], Susan R. Berman[4]

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 and discrete achievement levels on assessments used for accountability purposes. The present work scales up and extends predictive models of math test scores and achievement levels from existing literature and specifies six categories of models that incorporate information about student prior knowledge, demographics, and performance within the MATHia intelligent tutoring system. Linear regression, ordinal logistic regression, and random forest regression and classification 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.

Notes for Practice

- Advanced educational technologies, including simulations, games, and intelligent tutoring systems, continually assess students in order to provide them with appropriate activities and to determine their mastery of the topics presented.
- The assessment embedded in adaptive systems is a type of formative assessment, but we can also use it to make summative conclusions about what a student has learned.
- We show that process data collected from students using MATHia, an intelligent tutoring system, over the course of a year can predict high-stakes test scores over and above the ability of a prior-year test to predict these scores.
- Models learned on data from a single academic year can be used to predict outcomes for students in other academic years, suggesting that significant predictors of student outcomes remain relatively stable from year to year.
- The ability to predict high-stakes exam scores is a necessary (though insufficient) step towards replacing such exams with embedded formative assessments, but even if high-stakes exams remain in place, predictive tools can provide important information about learner readiness for such high-stakes exams.

Keywords

Intelligent tutorial systems, formative assessment, mathematics education, accountability, assessment, predictive modelling

Submitted: 14.06.2018 — **Accepted:** 24.01.2019 — **Published:** 05.08.2019

Corresponding author [1] Email: ggzheng@uga.edu Address: Department of Educational Psychology University of Georgia, 110 Carlton Street, Athens, 30602, GA, USA

[2] Email: sfancsali@carnegielearning.com Address: Carnegie Learning, Inc., Union Trust Building, 501 Grant St #1075, Pittsburgh, PA 15219, USA

[3] Email: sritter@carnegielearning.com Address: Carnegie Learning, Inc., Union Trust Building, 501 Grant St #1075, Pittsburgh, PA 15219, USA

[4] Email: sberman@carnegielearning.com Address: Carnegie Learning, Inc., Union Trust Building, 501 Grant St #1075, Pittsburgh, PA 15219, USA

1. Introduction

Formative and summative assessments differ in their intentions. Black and Wiliam (1998) defined formative assessment as "encompassing all those activities undertaken by teachers, and/or by their students, which provide information to be used as feedback to modify the teaching and learning activities in which they are engaged." In contrast, summative assessments summarize student achievement to a particular point in time. They are often used for accountability purposes: to determine whether students, teachers, schools, or curricula are achieving to desired levels.

In addition to their differences in intent, they also tend to differ in format and environment (e.g., Harlen & James, 1997). Summative assessments are typically given as part of a defined "test time" in an environment designed to remove distractions and interactions with other students. This "test time" is intended to be different from instructional time (Heritage, 2010). In fact, statistical assumptions used in interpreting summative assessments assume that learning does not take place during the exam, and the environment is often structured in order to make that assumption valid. Formative assessments, in contrast, may be informal and not clearly distinguished from instruction. In practice, it is common for formative assessments to mimic the form of summative assessments (taking the form of a specific set of test items completed individually by students), usually with less formal proctoring, but this need not be the case. Wiliam (2011) argues that focusing on the form of a formative assessment is a mistake; instead, formative assessment should be thought of as a process for evaluating student knowledge in order to modify instruction, rather than the instrument used for that purpose.

Educational technologies have the potential to change the way we think about formative assessments and their relationship to summative conclusions about learning (Shute, Levy, Baker, Zapata, & Beck, 2009). Although some educational technologies do distinguish between instructional and assessment environments in a teach–assess–reteach cycle, other educational technologies — particularly simulations, games, and intelligent tutoring systems — fully integrate formative assessment into the instruction (Mislevy, Steinberg, & Almond, 2003). In a game, for example, students typically "level up" by reaching a particular level of proficiency in playing the game itself, rather than using gameplay as a way to learn and then later demonstrate proficiency in a test environment. The game is continually formatively assessing the student, and the result of this gameplay is a summative conclusion: that the student is ready to advance to a more difficult level (Shute & Ke, 2012). Similarly, many intelligent tutoring systems continually assess students for the purposes of picking activities that are best suited to individual students and for determining progress along a mastery sequence.

This opportunity to rethink the relationship between formative and summative assessment comes at a time of increasing concern about current approaches to summative assessment and school accountability systems (Perie, Marion, & Gong, 2009; Evans & Lyons, 2017). Some of these concerns reflect the disconnect between assessment and instruction (Brookhart, 2009). The public sees standardized tests as nothing like the kind of authentic, rich, problem-solving focused activities that they believe are at the heart of good instruction. For example, in a recent poll, only 6% of parents see standardized test outcomes as being reflective of school quality, as opposed to 36% who saw teaching "co-operation, respect and problem solving" as indicative of school quality (PDK/Gallup, 2015). In the same poll, parents ranked testing last as a focus of school improvement (behind teacher quality, setting high standards, principal quality, and funding), as a measure of school effectiveness (behind student engagement, student hopefulness, graduation rate, college enrollment, and employment after graduation) and student academic progress (behind grades, teacher observations, and examples of student work).

Summative testing environments that are distinct from instructional environments take time away from instruction (Lazarín, 2014). Large school districts recently surveyed by The Council of Great City Schools (Hart et al., 2015) reported that, over a typical academic year and for all subjects, the average eighth grader spent 25.3 hours taking 10.3 tests. This statistic only considers district-administered tests. The time taken for summative testing, when including teacher-administered tests, is certainly much larger. With respect to high-stakes exams, the time spent preparing for the exam is typically much larger than the time to take the test itself. Nelson (2013), for example, studied two midsize urban school districts. One district devoted 15 hours per year to district tests and 80 hours preparing for those tests (including only time spent taking practice tests and sessions devoted to test-taking techniques). The other spent 55 hours per year taking tests and over 100 hours on test preparation. It is hard to say how typical these results are over all schools, but combined test taking and preparation times in these districts' figures represent over 10% of school time. With respect to instruction in highly tested subjects like mathematics, the proportion of instructional time may be much higher. Based on administrative and teacher reports, 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. Schools in the district we study here are directed to administer 10 benchmark exams, each of which takes one class period, plus a class period for preparation and a period going over results. Ten additional days are taken for preparation and administration of the end-of-year exam. Thus, in this district, over 20% of instructional time in mathematics is spent on preparation, review, and administration of standardized tests.

Our focus on summative assessments taking time away from instruction is not meant to suggest that summative testing can have no role in improving student learning. Indeed, there is good evidence (Roediger & Karpicke, 2006) that testing can improve learning, even when no feedback is given. Such studies, however, emphasize the psychological effect of "testing" memory (recalling facts, rather than studying or rehearsing them). While formal summative assessments do require students to recall facts, they are not the only (and almost certainly not the most effective) techniques that require students to test their memory. Most active learning environments are also tests in this sense. Summative assessments rarely give feedback, for example, and, although many studies find that students will learn from testing themselves, even in the absence of feedback, most studies find that feedback enhances learning (Pashler, Cepeda, Wixted, & Rohrer, 2005). Our argument, then, is not that testing cannot be a valuable educational experience, but that summative testing is not designed to be educational. For this reason, they are likely less effective educational experiences than ones designed from the start to help students learn.

There are other characteristics of high-stakes summative assessments that contradict what we know about how people learn (Snow & Lohman, 1989). For example, such tests inherently emphasize the importance of a student's knowledge at one particular time: the day of the test. The implicit assumption here is that maximizing performance on that day equates to maximizing what we really care about (performance in the long term). This emphasis on peak knowledge encourages cramming for the test, as opposed to distributing practice over a period of time. Concentrated studying produces increases in short-term recall, thus helping with immediate performance on the test, but this practice schedule will tend to decrease long-term retention (Carpenter, Cepeda, Rohrer, Kang, & Pashler, 2012). Thus, the assumption underlying the high-stakes end-of-course exam is simply not true. Optimizing knowledge for a single point in time is not the same as optimizing knowledge for long-term retention.

A related problem with high-stakes testing is that it presumes that students learn at approximately the same rate. In typical school environments, instructional time is (relatively) fixed. A course has a specific number of instructional hours (the main time variable is the amount of out-of-class homework and studying time). The student's job is to learn the required amount of that material by the day of the test. But we know that, for various reasons, students will take differing amounts of time to learn material. Gettinger and White (1980) found learning rates between students differed by up to a factor of four. Zerr et al. (2018) also found strong differences in student learning efficiency. One of the barriers to implementing competency-based progression is that the assessments required to progress are so infrequent. An assessment system that adjusts to the student's rate of learning could enable progression at any time.

The public's lack of faith in testing, combined with the US education system's fairly recent emphasis on formal summative assessments as a major component of their accountability systems, has led to an urgent focus on accountability systems that rely less on high-stakes testing. In New York State (and nine other states in the US), parents are legally entitled to have their students forgo some high-stakes exams. In the 2016–2017 school year, 27% of New York parents opted their students out of high-stakes testing in math (Moses, 2017). So many students in Minneapolis recently opted-out of state exams for 10th and 11th grade math that the state does not recommend relying on exam results (State of Minnesota, 2017). States like Maryland and New York have responded by mandating reduced testing time (Walker, 2018), and Georgia has created a pilot program focused on reducing summative testing time (Tagami, 2018).

Recognition of the problems with standardized testing is not new, but the advance of educational technology holds new promise to change the way that summative testing gets done. Twenty years ago, Grigorenko and Sternberg (1998) 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 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. (Grigorenko & Sternberg, 1998)

Further, they point out historical antecedents for approaches that assess students while they learn going back to at least the early 20th century. For example, they point to Binet (1909) as both the creator of what they call static testing and as an advocate of process assessment. Grigorenko and Sternberg (1998) summarize Buckingham's (1921) view 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."

Various recent approaches take seriously what Campione and Brown (1985) 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 are able to demonstrate that this instruction-embedded formative assessment can provide us with information about student abilities equivalent to that provided by high-stakes summative assessments, then maybe we can start to replace summative assessments with continuous adaptive instruction.

As a first step, demonstrating the validity and reliability of such predictive models is necessary. 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

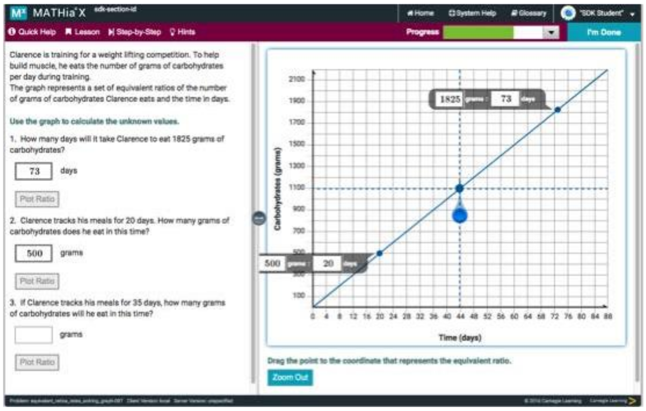
(Ritter, Anderson, Koedinger, & Corbett, 2007), on which the MATHia platform is based. Next, we explain the dataset for the present study and our model specification approach. Finally, we provide our results, discuss these results as well as their limitations, and present avenues for future research in predictive modelling 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 ongoing 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, sociodemographic 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 technology (Ritter et al., 2007), 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 (Pane, Griffin, McCaffrey, & Karam, 2014). Carnegie Learning's blended model calls for a 60%–40% split between student-centred, 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 (Bloom, 1968); in each workspace, there are multiple, fine-grained knowledge components (KCs or skills; Koedinger, Corbett, & Perfetti, 2012) 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; Corbett & Anderson, 1994).



Clarence is training for a weight lifting competition. To help build muscle, he eats the number of grams of carbohydrates per day during training.
The graph represents a set of equivalent ratios of the number of grams of carbohydrates Clarence eats and the time in days.

Use the graph to calculate the unknown values.

1. How many days will it take Clarence to eat 1825 grams of carbohydrates?

73 days

2. Clarence tracks his meals for 20 days. How many grams of carbohydrates does he eat in this time?

500 grams

3. If Clarence tracks his meals for 35 days, how many grams of carbohydrates will he eat in this time?

grams

Drag the point to the coordinate that represents the equivalent ratio.

Figure 1: Screenshot from Carnegie Learning's MATHia intelligent tutoring system, based on Cognitive Tutor technology (©2019 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 to create predictive models for innovative assessment in domains outside of mathematics have been pursued in domains like reading (e.g., Beck, Lia, & Mostow, 2004) and physics (e.g., Shute & Moore, 2017), but even a brief review of such work is beyond the scope of this paper.

3.1. Using ASSISTments Data

Numerous papers (Anozie & Junker, 2006; Ayers & Junker, 2008; Feng, Heffernan, & Koedinger, 2006; Junker, 2006; Pardos, Heffernan, Anderson, Heffernan, & Schools, 2010) have considered various elements of process data (e.g., percent correct on various types of items, metrics like hint-seeking, number of KCs mastered) from the ASSISTments system (Razzaq et al., 2005) 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 (Pardos, Baker, San Pedro, Gowda, & Gowda, 2014) relying on the predictions of so-called "detector" models of behaviour and affective states (Baker et al., 2012).

Work by Feng et al. (2006) 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 learner's 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, Feng et al. (2006) report a within-sample MAD of 5.533 points over a sample of 600 ASSISTments learners in 2004–2005. This MAD represents an error rate of 10.25% 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. (2014) report both MAD and correlations of their predicted values on the MCAS to learners' actual MCAS scores using models that incorporate elements of learner behaviour (e.g., gaming the system; Baker, Corbett, Roll, & Koedinger, 2008), affective states (e.g., boredom; Baker et al., 2012), 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-labelling schema follows that provided by Ritter, Joshi, Fancsali, and Nixon (2013), 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 Standards of Learning (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 Ritter et al. (2013) are a subset of variables we consider in this work. 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; Schwarz, 1978) decreased, indicating a justification for increasing the complexity of the predictive models because of increases in predictive performance achieved.

Table 1: Variable Sets Considered in This Work

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

Note: Pre-test = pre-test score; process = process variables from MATHia usage; demog = demographic variables); M1–M5 are also considered by Ritter et al. (2013); all sets include learner grade-level (6–8).

Table 2: Summary of Model Fits for Predicting Virginia SOL in Grade 7

Model	# Vars	BIC	R^2
M1 (pre-test)	1	2,041.5	0.50
M2 (process)	5	2,181.0	0.43
M3 (process + demog)	7	2,167.8	0.45
M4 (pre-test + demog)	3	2,030.6	0.51
M5 (pre-test + demog + process)	8	1,928.4	0.57

Note: Data based on different sets of variables, reproduced from Table 4 in Ritter et al. (2013).

Ritter et al. (2013) 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 predictiveness. 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 of test scores 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 Ritter et al., 2013).

Later work by Joshi, Fancsali, Ritter, Nixon, and Berman (2014) 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 at levels similar to those achieved by previous efforts (Pardos et al., 2014; Ritter et al., 2013; R^2 of the best model ~0.32).

Overall, models reported in Table 2 for the Virginia SOL compare favourably with models reported by Pardos et al. (2014) 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, interpreting the R^2 of these predictive models as the square of observed correlations between predicted and actual MCAS scores.

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 predictive performance.

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 the fourth largest school district in the United States (US Department of Education, 2016).

In each year, the school district provided, for each student, grade-level (i.e., 6–8), current year standardized test scores, previous year standardized test scores (pre-test), and demographic data that could then be mapped to usage data from the MATHia system. In 2013–2014, 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 modelling, but we report statistical accuracy in terms of more interpretable FCAT and FSA scale score units.

Table 3: Sample Sizes by Grade-Level (Grade) and Academic Year (13–14 = 2013–2014, etc.)

Grade	13–14	14–15	15–16	Total
6	2,914.0	2,471.0	3,542.0	8,927.0
7	3,827.0	3,596.0	3,505.0	10,928.0
8	1,200.0	1,301.0	1,018.0	3,519.0
All	7,941.0	7,368.0	8,065.0	23,374.0

The FCAT exam provides a developmental scale score ranging from 140 to 298 from Grades 3 to 8 (Florida Department of Education, 2014), and the FSA exam provides a developmental scale score ranging from 240 to 393 from Grades 3 to 8 (Florida Department of Education, 2017). 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.

Demographic variables (demog) considered (and most frequently occurring values) include:

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

We consider process variables drawn from the set of variables considered by previous work with Cognitive Tutor predicting Virginia’s SOL exam (Ritter et al., 2013) as well as several novel variables. As in previous work (e.g. Figure 2 in Ritter et al., 2013), 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 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 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, and Results

As noted, we consider pre-test scores (pre-test) and categories of process variables (process) and 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 three methods of specifying and learning non-hierarchical regression models to predict standardized test scores as well as two methods for specifying and learning non-hierarchical classification models to predict standardized test achievement levels. We consider both regression and classification results before considering 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 regression models that account for the inherent hierarchy of this type of data (e.g., students working within schools) versus the regression models in this section that do not explicitly do so.[1]

5.1. Non-Hierarchical Models: Regression

Despite the previous success of relatively simple, linear regression approaches to this problem (e.g., Feng et al., 2006; Ritter et al., 2013), we compare three approaches to regression: ridge regression (Hoerl & Kennard, 1970), stepwise (ordinary least squares) linear regression, and random forest regression (Breiman, 2001). Ridge regression and random forest regression represent more sophisticated approaches compared to relatively simple ordinary least squares, stepwise regression. Ridge regression is often used in cases in which there are many candidate predictors (as will especially be the case as we consider interaction terms and quadratic transformations among possible predictors), and many of these predictors may be correlated with each other. Two-way interaction terms are important to include as candidate variables. Consider, for example, the case of students with exceptional status like those with a learning disability or with "gifted" status. While students with one or more learning disabilities may require more time to complete content, gifted students may take less time, and these differences may manifest in models in which two-way interaction terms are statistically significant and contribute to better predictions. Including quadratic terms of continuous process variables in the set of candidate variables allows us to consider one possible form of non-linear dependence between process variables and high-stakes test performance, indicated by several scatterplots considered in initial exploratory analysis of data. Further, interaction and quadratic terms contribute to maximizing possible predictive accuracy of models we consider.

In ridge regression, coefficients for predictors are decreased or "regularized." This approach decreases the extent to which models over-fit the data on which they are trained. The random forest regression approach is likely to be appropriate both in cases in which there are a large number of candidate predictor variables as well as when linearity and parametric assumptions such as normality fail in data being considered.

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

5.1.1. Ridge Regression

Ridge regression (Hoerl & Kennard, 1970) is a regularized form of linear regression that accounts for multi-collinearity inherent to large sets of features (i.e., correlations among predictor variables) like those considered by the present work. The approach "regularizes" (biased) slope coefficients in a regression model to lower values (though not to zero as in the related least absolute shrinkage and selection operator or LASSO regression; Tibshirani, 1996[2]), decreasing the extent to which models are likely over-fit on training data. Suppose there are p slope coefficients in the model and n elements in the data set, estimating a ridge regression model minimizes the sum of residual sum of squares with a regularization term:

$$\sum_{i=1}^{n}(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij})^2+\lambda\sum_{j=1}^{p}\beta_j^2 \text{ (Eq. 1),}$$

where λ is a tuning parameter for the regularization term, which is selected by cross validation. We fit ridge regression models using the maximal set of model variables for each of the six candidate sets of variables (M1 through M6). Categorical variables such as demographic variables are all dummy coded before ridge regression models are fit.

5.1.2. Stepwise Linear Regression (SLR)

In contrast to ridge regression, we also use a stepwise procedure to find the best ordinary least squares regression model for six candidate sets of predictors (i.e., for 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 as a model selection criterion provides for simpler models in terms of the number of variables that will be included, while also generally providing for better generalizability for held out data sets (Raftery, 1995).

5.1.3. Random Forest Regression (RFR)

Random forest models (Breiman, 2001) are learned by an ensemble machine learning method that pairs decision tree learning with bootstrap aggregation ("bagging"; Breiman, 1996) to produce models that are unbiased and 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 (RFR). 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 fewer than five cases.

5.1.4. Reporting Statistical Accuracy

We consider the accuracy of the models that predict scale scores selected by the procedures above for the variable sets corresponding to M1 through M6 on training data from the 2013–2014 school year in terms of their mean absolute deviation (MAD) and root mean square error (RMSE) in predicting learners' FCAT and FSA scores in terms of raw scale score points:

$$MAD=\frac{1}{n}\sum_{i=0}^{n}\left|FSA_i-\widehat{FSA}_i\right| \text{ (Eq. 2)}$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=0}^{n}(FSA_i-\widehat{FSA}_i)^2} \text{ (Eq. 3),}$$

where n is the sample size; $\widehat{FSA}_i$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 Equations 2 and 3 when testing on 2013–2014 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 the R^2 values of the models.

[2] This is an important feature of ridge regression in the present context since we include both interaction and quadratic terms in our predictive models. Without imposing further constraints on model estimation, a LASSO approach may shrink coefficients of main terms to zero while estimating non-zero coefficients for corresponding interaction and quadratic terms. Our stepwise linear regression approach is constrained to "keep" main terms if their corresponding interaction and/or quadratic terms are elected in the model.

5.1.5. Questions for Analysis

In the context of regression models, we present analytical results addressing the following three questions:

- **RQ1**: What are the relative performance characteristics of RIDGE, SLR, and RFR models in predicting test scores when models are learned over variable sets M1 through M6?
- **RQ2**: What accuracy can be achieved?
- **RQ3**: What are the relative contributions of different categories of variables included in these models?

We focus discussion on models trained on data from the 2013–2014 and 2014–2015 school years, testing such models on data from the other two years from which data were available. Results suggest consistent, broad qualitative patterns when using 2013–2014 as training data and using 2014–2015 as training data, including approximate accuracy achieved, differences in accuracy over models M1 through M6, corresponding differences in BIC scores indicating statistically significant differences in models according to heuristics due to Raftery (1995), differences between MAD and RMSE on the training and test data sets, etc. This is perhaps surprising given that 2014–2015 data reflected more usage for the typical student in general than 2013–2014 data, and the FCAT was used in 2013–2014 while the FSA was used in subsequent years. Further, statistical tests comparing performance of these methods over the four held-out data sets on which trained models were tested suggest a class of models that are statistically indistinguishable.

5.1.6. Regression Results

Tables 4 and 5 provide RMSE and MAD, respectively, as measures of accuracy for the RIDGE, SLR, and RFR models learned on 2013–2014 and applied to data from 2014–2015 and 2015–2016. Tables 6 and 7 provide these accuracy measures for models trained on 2014–2015 data and tested on 2013–2014 and 2015–2016 data. On both measures and across all four test data sets, RIDGE and SLR outperform RFR in all but two cases; RFR outperforms SLR (but not RIDGE) for M5 learned on 2014-15 data with 2013–2014 test data in Table 4 and for M4 learned on 2014-15 data for 2015–2016 test data in Table 6. We also see that accuracy tends to improve in the progression of M1 to M6, but the size of differences along both dimensions (i.e., between RIDGE, SLR, and RFR models and between models M1–M6) are relatively small.

To more rigorously address RQ1, we follow an approach recently advocated by Gardner and Brooks (2018) in this journal and by others (Demšar, 2006; García, Fernández, Luengo, & Herrera, 2010) to determine whether there are significant differences among these models and, perhaps more importantly, whether there are classes of models that are statistically indistinguishable. To do so, we rely on the Friedman Aligned Ranks method and a corresponding post hoc test (Hodges & Lehmann, 1962; García et al., 2010), an alternative to the Friedman test (Friedman, 1940) used by Gardner and Brooks (2018) that relies on more information than the latter test by making inter-dataset comparisons and by using information about performance of each algorithm relative the average performance achieved on a dataset by all other algorithms. Following Gardner and Brooks (2018), we consider each model M1 through M6 paired with each of RIDGE, SLR, and RFR as an algorithm (for a total of 18 algorithms) and the four test datasets across training regimes in Tables 4 and 6 (and Tables 5 and 7, as we run the test using RMSE and MAD performance metrics separately). Later, we apply the same statistical test regime to the classification task in §5.2.[3]

Using the Friedman Aligned Ranks test, we find statistically significant differences between the eighteen candidate algorithms over the four datasets when considering both RMSE and MAD (test statistic T = 64.762, $p < .00001$ for RMSE; T = 64.273, $p < .00001$ for MAD). While this conclusion by itself is not especially informative, post hoc tests allow us to make pairwise comparisons to determine whether some models perform in a way that is indistinguishable from the model with the greatest accuracy. We use the post hoc procedure suggested for this approach by García et al. (2010), which is the same as that of the Kruskal-Wallis test (1952). From the value of this statistic, we can determine a p-value that can be compared to an

[3] Our training and testing scheme deviates from those deployed in the existing literature that uses this type of evaluation procedure (Gardner & Brooks, 2018; Demšar, 2006; García et al., 2010) in at least two important ways: 1) the two 2015–2016 test datasets (out of four) are not independent but rather identical; we train models on different datasets and test them against the same dataset as well as two independent datasets, and 2) existing studies focus on cases in which performance is measured in a cross validation scheme with multiple test folds. Cross-validation naturally returns a sampling distribution of accuracy measures given that each fold is a random sample from the underlying population, and such sampling distributions may be used to generalize the results to other test sets that are from the same population as the cross-validation data set. With independent and fixed training and test datasets, as in the present case, there are fixed measures of accuracy on these test datasets and no sampling distribution from which to derive an accuracy measure, but we assume the measure is appropriately "reliable." To address 1), we checked robustness for the regression case and the RMSE accuracy measure by dropping one of the non-independent datasets from the set of testing datasets, and the conclusions we reach were identical to the case in which all four testing datasets were considered. Given numerous options for evaluating these kinds of models, including those of the frequentist and Bayesian variety (Gardner & Brooks, 2018), we take it to be an open question which evaluation methods are most appropriate in this case and other cases.

appropriately adjusted p-value to determine whether any pair of algorithms exhibits a statistically significant difference in performance. We rely on the simple Bonferroni-Dunn procedure for 153 pairwise comparisons (18 choose 2).

The comparisons we primarily care about involve determining which models are not significantly different in performance from the model(s) with the best-observed accuracy measure. Model M5 with SLR exhibits the best average aligned rank in terms of its RMSE performance (M5 + RIDGE receives this honour on MAD results), and post hoc testing on RMSE results indicates that the M5 + SLR model is indistinguishable from the other two M5 models as well as all three M6 models. The results of the same testing regime on MAD results yield the same class of indistinguishable models. Thus, the class of M5 and M6 models out-perform those of M1–M4 but are indistinguishable from each other. Given this inability to distinguish between M5 and M6 models and the relative simplicity of SLR, we focus on the results of SLR.

With respect to the accuracy our models achieve (RQ2), 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 (2013–2014) as well as on the two test sets (2014–2015 and 2015–2016).

Table 4: Comparison of the Accuracy of Models Learned by RIDGE, SLR, and RFR, expressed as RMSE

	14–15			15–16		
Model	RIDGE	SLR	RFR	RIDGE	SLR	RFR
M1	12.730	12.726	13.483	12.955	12.968	13.716
M2	13.688	13.640	13.872	13.948	13.983	14.431
M3	13.267	13.161	13.390	13.388	13.384	13.822
M4	12.066	12.328	12.487	12.264	12.657	12.702
M5	11.236	11.212	11.282	11.303	11.179	11.508
M6	11.324	11.278	11.447	11.400	11.256	11.658

Note: For models learned on 2013–2014 data applied to data from the 2014–2015 (14–15) and 2015–2016 (15–16) school year.

Table 5: Comparison of the Accuracy of Models Learned by RIDGE, SLR, and RFR, expressed as MAD

	14-15			15-16		
Model	RIDGE	SLR	RFR	RIDGE	SLR	RFR
M1	9.883	9.888	10.416	10.017	10.023	10.524
M2	10.539	10.495	10.722	10.693	10.722	11.110
M3	10.191	10.142	10.327	10.222	10.249	10.604
M4	9.356	9.541	9.675	9.435	9.729	9.789
M5	8.657	8.669	8.717	8.696	8.597	8.914
M6	8.753	8.736	8.847	8.782	8.673	9.028

Note: For models learned on 2013–2014 data applied to data from the 2014–2015 (14–15) and 2015–2016 (15–16) school year.

To qualitatively compare these modelling results to those in the previous literature on predicting standardized test scores from similar types of data, we provide test data R^2 values for the best performing SLR models (in terms of BIC score) in Table 8. These R^2 values, especially for M5 and M6, are considerably larger than those reported in prior literature.[4]

As noted, Figure 2 provides scatterplots of predicted values of FCAT and FSA scores against learners' actual or "true" values for these models.[5]

[4] Notably, R^2 values were not used as a model selection criterion; they are only provided as a way of comparing these results to those from prior literature.

[5] A reviewer points out that Figure 2 indicates that the residuals of our regression models are not strictly homoscedastic, especially due to individuals who 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 who are likely to perform especially

Table 6: Comparison of the Accuracy of Models
Learned by RIDGE, SLR, and RFR, expressed as RMSE

	13-14			15-16		
Model	RIDGE	SLR	RFR	RIDGE	SLR	RFR
M1	11.852	11.837	12.658	12.833	12.830	13.513
M2	12.621	12.601	12.964	14.024	13.955	14.287
M3	12.466	12.462	12.626	13.585	13.633	13.807
M4	11.577	11.780	12.000	12.152	12.562	12.494
M5	10.838	10.893	10.954	11.432	11.436	11.620
M6	10.746	10.749	10.996	11.524	11.472	11.694

Note: For models learned on 2014–2015 data applied to data from the 2013–2014 (13–14) and 2015–2016 (15–16) school year

Table 7: Comparison of the Accuracy of Models
Learned by RIDGE, SLR, and RFR, expressed as MAD

	13-14			15-16		
Model	RIDGE	SLR	RFR	RIDGE	SLR	RFR
M1	8.933	8.923	9.455	9.940	9.927	10.434
M2	9.596	9.572	9.879	10.866	10.805	11.126
M3	9.611	9.641	9.688	10.526	10.595	10.752
M4	8.914	8.970	9.170	9.370	9.666	9.687
M5	8.289	8.396	8.376	8.834	8.846	9.054
M6	8.152	8.215	8.373	8.923	8.919	9.122

Note: For models learned on 2014–2015 data applied to data from the 2013–2014 (13–14) and 2015–2016 (15–16) school year.

Table 8: R^2 Values for SLR Models
Learned on 2013–2014 Data (Training: 13-14) and 2014-2015 Data (Training: 14-15)

	Training: 13-14		Training: 14-15	
Model	14–15	15–16	13–14	15–16
M1	0.603	0.642	0.601	0.650
M2	0.544	0.584	0.548	0.586
M3	0.575	0.619	0.558	0.605
M4	0.627	0.659	0.605	0.664
M5	0.692	0.734	0.662	0.723
M6	0.688	0.730	0.671	0.720

Note: Applied to test data from 2014–2015 (14–15) and 2015–2016 (15–16) and applied to test data from 2013–2014 (13–14) and 2015–2016 (15–16), respectively; while not used in model selection, these values provide for comparisons to models reported in Ritter et al. (2013) and Pardos et al. (2014).

To address RQ3, we find that relying on pre-test data alone (M1) provides for predictions that are approximately one point better (though sometimes less and in some cases slightly more) than relying on process data alone (M2), whether models are trained on 2013–2014 data or 2014–2015 data (see Tables 4–7), but post hoc tests for pairwise comparisons indicate that the models are mostly indistinguishable from each other. Each of these "minimal" models (only pre-test or only

poorly (or possibly exceptionally well). If so, such models might be useful for targeting intensive intervention.

process data) can account for variability in held-out test data at a level comparable to the best models reported by Ritter et al. (2013) on training data and to the best models for ASSISTments tested on held-out data reported by Pardos et al. (2014).

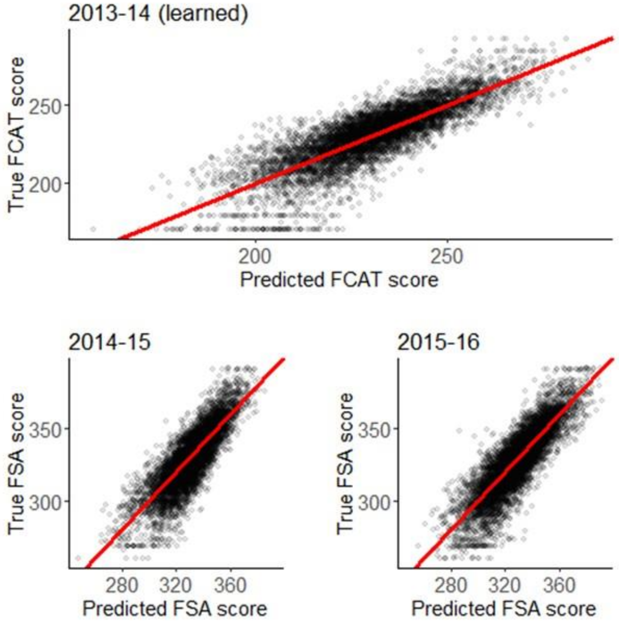


Figure 2: Scatterplots of predicted values of FCAT and FSA scores against actual scores from M5 learned using SLR on 2013–2014 data and applied to test data from 2014–2015 and 2015–2016**.** The red line is a reference line from the graph origin with slope = 1. Table 9 presents details of this model. FCAT and FSA scores are on different scales, but our predictive models produce z-scores, which are then transformed to arrive at an appropriate scale score prediction.

To address RQ3, we find that relying on pre-test data alone (M1) provides for predictions that are approximately one point better (though sometimes less and in some cases slightly more) than relying on process data alone (M2), whether models are trained on 2013–2014 data or 2014–2015 data (see Tables 4–7), but post hoc tests for pairwise comparisons indicate that the models are mostly indistinguishable from each other. Each of these "minimal" models (only pre-test or only process data) can account for variability in held-out test data at a level comparable to the best models reported by Ritter et al. (2013) on training data and to the best models for ASSISTments tested on held-out data reported by Pardos et al. (2014).

Adding demographic data to each of these provides for modest improvements in accuracy, but models that include pre-test, process, and demographic (demog) variables (M5) are found to be in the class of models that perform best on these data, along with M6 models that drop demographic data, indicating that demographic data may not be necessary to achieve satisfactory models.

This points to an important opportunity in future work. Not relying on demographic data for predicting test scores is important from the standpoint of practical and ethical implementations of these predictive models. We believe it is unacceptable for a system to handicap (or boost) a student's predicted score simply because of their status with respect to certain demographic categories (e.g., ethnicity). To do so could "perpetuate the biases and prejudices in cultural, geopolitical, economic and societal realities" (Slade & Prinsloo, 2013). At the very least, including demographic data is likely to raise serious barriers to adoption of learning platforms, and such data are often not available to learning platform vendors at any reasonable scale in any event. In addition to concerns of explicit bias that could be introduced by relying on such demographic categories, concerns of other forms of algorithmic bias (e.g., Danks & London, 2017; O'Neil, 2016) 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. In practical use cases in which predictive analytics will be provided to teachers and other stakeholders at scale, it is advantageous to be able to rely on models like M2 that only include student process and performance data for the current year, as these data are directly available from student usage of a system like MATHia. Other data like demographics and pre-test scores are not generally readily available to a system like MATHia.

Insights from the demog, process, and pre-test variables found to be significant in M5 are worth briefly considering. We provide a summary of the regression model M5, including parameter estimates and significance, learned by SLR in Table 9, though we emphasize that our primary interpretive goal in this work is to paint a relatively "big" picture of the variables contributing to the model, rather than provide strong specific interpretations of each individual predictor and its significance (or lack thereof).

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. (2013). These models would appear to illuminate a different, though related, set of important indicators of progress within the MATHia system compared to this previous work.

Also consistent with existing literature (e.g., Sirin, 2005), we find that a variety of demographic variables are significant predictors of exam score. While variables that encode ethnicity are 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).

Table 9: Estimated SLR Model M5 Learned on 2013–2014 Data

Variable	Coefficient
Intercept	-1.141***
Pre-test	0.488***
Grade 7	0.342***
Grade 8	0.544***
Total problem solving time	-0.250***
Problems per workspace	0.304***
Number of KCs mastered	0.179***
Problems per workspace x Number of KCs mastered	-0.134***
Pre-test x Problems per workspace	0.041***
Number of workspaces encountered	0.247***
ESE: Gifted	0.198***
ESE: Learning disability	-0.024
ESE: Other	-0.002
LEP: Enrolled	-0.103
LEP: Former	0.021
LEP: Not Enrolled	-0.023

Note: Standardized parameter estimates; results of this model applied to data from 2014–2015 and 2015–2016 school years reported in Tables 4–5; ***p < 0.001; **p < 0.01; *p < 0.05.

The significance of ESE and LEP status indicators may provide an important pointer to areas for future focus. That these indicators are significant predictors of the exam score indicates other possible process variables that describe student progress within MATHia but are not yet included in our models. In the future, identifying and including these variables in these types of models can help override the predictability of demographic variables such as ESE and LEP, leading us to rely less on such status indicators (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.

5.2. Non-Hierarchical Models: Classification

Recall that the FCAT and FSA standardized tests have both scale scores and discrete achievement levels from 1 to 5, with 1 being the lowest score and 5 the highest score (and a score greater than or equal to 3 counts as "passing"). We now consider the problem of predicting student achievement levels on these standardized tests, adopting an approach similar to that adopted in the previous section, relying on stepwise ordinal logistic regression models as well as random forest classification (or classifier) models and the same statistical testing regime to compare models that result.

5.2.1. Stepwise Ordinal Logistic Regression (SOLR)

Ordinal logistic regression is a generalized linear model used to model relationships between predictor variables and an ordinal response variable. In data sets we consider, the FCAT and FSA exams have five naturally ordered achievement levels. Suppose we have predictor variables, *X1, X2, …, Xs*. Ordinal logistic regression assumes that the log odds of levels at or below level k (in this case k takes integer values from 1 to 5) versus levels above k can be written as a linear function of the predictor variables. It also assumes that only the intercept *αk* varies with level k, and all the slopes in the model do not (Equation 4):

$$\log\left(\frac{P(Y\leq k)}{P(Y>k)}\right) = \alpha_k + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_s X_s, k \in \{1,2,3,4,5\} \text{ (Eq. 4)}$$

We run (stepwise) ordinal logistic regression (SOLR) using the R function *poly* in the *MASS* package (Venables & Ripley, 2002).

5.2.2. Random Forest Classifier (RFC)

In contrast to the RFR approach adopted in §5.1.3, we adopt a random forest classifier approach to predict ordinal achievement levels on the FCAT and FSA. Similar to regression trees that are components of RFR models, classification trees divide the predictor space into a set of hyper-rectangles. However, the cutting or split rule is to maximize the purity (or alternatively, information gain) at each node of the tree. The prediction is determined for each node by choosing the most commonly occurring achievement level at that node.

5.2.3. Reporting Classification Accuracy

In considering performance of models to predict achievement levels on the FSA and FCAT exams, we denote actual FSA achievement level for student i as $FSAL_i$ (or $FCATL_i$ for the FCAT for student i achievement level on the FCAT in 2013–2014) and denote predicted achievement levels $\widehat{FSAL_\iota}$ (or $\widehat{FCATL_\iota}$ in 2013–2014). We report misclassification rate (MR; Equation 5) and define the off-by-one rate (OBOR; Equation 6):

$$MR = \frac{1}{n}\sum_{i=1}^{n} I(FSAL_i \neq \widehat{FSAL_i}) \text{ (Eq. 5)}$$

$$OBOR = \frac{1}{n}\sum_{i=1}^{n} I(|FSAL_i - \widehat{FSAL_i}| \leq 1) \text{ (Eq. 6),}$$

where n is the sample size of the data set, and I is the indicator function that takes the value 1 when its argument is true. Both of these measures are more appropriate in the context of the goal of classification of achievement level, providing a measure of the predictive success we achieve in getting the achievement level (in)correct and the extent to which we are (at most) off-by-one achievement level in our predictions. The statistical performance measures used in regression analyses in the previous section, specifically MAD and RMSE, can also be used in developing classification models, but we omit extensive reporting of them for the sake of brevity. This omission is reasonable because patterns in the performance of classification approaches we consider are consistent across MR, OBOR, MAD, and RMSE.

5.2.4. Questions for Analysis

In the context of classification models, we present our results around providing answers to the following three questions:

- **RQ4**: What are relative performance characteristics of SOLR models and RFC models in classifying achievement levels?
- **RQ5**: What accuracy can be achieved?
- **RQ6**: What are the relative contributions of different categories of variables included in these models?

As in §5.1.5 and §5.1.6, we consider models trained on data from the 2013–2014 school year and on data from the 2014–2015 school year, using data from the other two years for testing; results suggest no significant deviations in the pattern of results found using 2013–2014 as training data when considering models learned using 2014–2015 training data when our predictive goal is classification instead of regression. This includes the results of Friedman aligned rank statistical testing indicating that the class of M5 and M6 models are statistically indistinguishable from one another and represent better accuracy than M1–M4 models when considering either of MR or OBOR accuracy.

5.2.5. Classification Results

Table 10 provides MR and OBOR measures for the accuracy of models that predict achievement levels on test sets from the years 2014–2015 and 2015–2016, with the training set from 2013–2014. We note that, unlike other accuracy measures, larger OBOR values indicate better performance. For all measures (i.e., RMSE, MAD, MR, and OBOR), SOLR outperforms RF, except in the case of model M1, which includes only pre-test data. Carrying out the same statistical testing regime we used for regression results (Friedman aligned rank test + post hoc testing for pairwise comparisons), in this case with twelve algorithms over four datasets, we again find that the class of M5 and M6 models are statistically indistinguishable and significantly different than M1–M4 models.

Table 10: SOLR and RFC Results for Models Learned on 2013–2014 Data Applied to Test Data from 2014–2015 (14–15) and 2015–2016 (15–16), Expressed as MR and OBOR

	MR				OBOR			
	14-15		15-16		14-15		15-16	
Model	SOLR	RFC	SOLR	RFC	SOLR	RFC	SOLR	RFC
M1	0.504	0.503	0.480	0.480	0.933	0.933	0.945	0.944
M2	0.524	0.553	0.509	0.541	0.936	0.913	0.933	0.911
M3	0.489	0.531	0.502	0.523	0.943	0.927	0.941	0.925
M4	0.492	0.496	0.472	0.478	0.941	0.934	0.950	0.946
M5	0.547	0.471	0.436	0.459	0.964	0.954	0.970	0.953
M6	0.459	0.477	0.443	0.464	0.964	0.950	0.967	0.952

Table 11: SOLR and RFC Results for Models Learned on 2014–2015 Data Applied to Test Data from 2013–2014 (13–14) and 2015–2016 (15–16), Expressed as MR and OBOR

	MR				OBOR			
	13-14		15-16		13-14		15-16	
Model	SOLR	RFC	SOLR	RFC	SOLR	RFC	SOLR	RFC
M1	0.500	0.490	0.503	0.486	0.925	0.950	0.926	0.949
M2	0.506	0.534	0.512	0.537	0.933	0.905	0.923	0.901
M3	0.492	0.522	0.484	0.514	0.937	0.919	0.932	0.913
M4	0.504	0.515	0.468	0.476	0.936	0.923	0.944	0.936
M5	0.458	0.468	0.431	0.448	0.956	0.947	0.963	0.953
M6	0.443	0.473	0.435	0.458	0.960	0.947	0.964	0.949

Considering performance on held-out test sets in Tables 10–11, the SOLR MR measure ranges from approximately 43% to almost 53%. However, inspecting RMSE and MAD, we find values that range from 0.75 to 0.86 and from 0.47 to 0.61, respectively, which indicates that, on average, the deviation between predicted and actual achievement level is less than one level. Rather than extensively report on such RMSE and MAD values, we note that these values are consistent with values of OBOR we observe for SOLR, all of which are greater than 0.92 (and greater than 0.90 for RFC); that is, at least 90–92% of the students in the test sets are predicted to be within one level of their actual level. While we recognize that there are important instances in which being off-by-one represents a failure to accurately predict whether a student will pass or fail the standardized test (e.g., when the model predicts 3, a passing achievement level, but the student actually achieves a level of 2, a failing achievement level), we take this to be a promising indicator of the success of these models in the classification context. More sophisticated classification regimes and models can also be used to develop models that specifically seek to avoid making such crucial mistakes, though this remains a topic for future research.

Results of the Friedman aligned rank test indicate that demographic variables may not provide significant value in helping to predict student achievement levels, as M5 and M6 models are not significantly different when compared in post hoc pairwise tests. M5 and M6 models both include pre-test and process variables, indicating that an indicator of students' prior knowledge (i.e., pre-test) and process variables together provide sufficient information about student achievement, providing what is likely a sufficient substitute for whatever information concerning student achievement is provided by demographic characteristics.

Carefully considering the results of M5, we report the importance of the predictors that contribute to these models, similar to our inspection of a particular M5 SLR model in Table 9. Table 12 reports the relative importance of each predictor variable measured by the mean decrease in the Gini index (MDG) of each predictor variable when constructing M5 with RFC.[6] The Gini "impurity" index (G) is calculated as potential "splits" (or "cuts") are considered that in turn define the

[6] Even though SOLR tends to result in better prediction accuracy than RFC, RFC generally provides for similar patterns in performance across the six candidate models, and M5 is still the top-performer among RFC models. Since the validity of p-values in ordinal logistic regression is dubious (Venables & Ripley, 2002), we focus on interpreting the predictor variables in M5 based on their MDG in constructing RFC models.

"decision nodes" (and corresponding hyper-rectangles or regions) of a decision tree. At each cut of the predictor space over a given predictor variable, the decrease in Gini index (G; Equation 7) before and after the cut is calculated:

$$G = \sum_{k=1}^{K} \sum_{m=1}^{M} \hat{p}_{mk}(1 - \hat{p}_{mk}) \text{ (Eq. 7),}$$

where $\hat{p}_{mk}$ is the misclassification rate of level k (i.e., in this case 1–5) in the m^{th} hyper-rectangle/region defined by the current cuts that have been conducted (or may be conducted). Smaller Gini indices indicate higher node purity, which means that a particular cut provides greater information about classification with respect to the target outcome. Decreases in Gini index are summed up across all the cuts made for this predictor. These summations are averaged across all the classification trees to obtain the mean decrease in Gini for each predictor, providing a measure of the importance of a particular predictor variable in terms of their contribution to a RFC model.

In Table 12, we observe that three process variables, including Problems per workspace, Workspaces mastered per hour, and Assistance per problem, outperform Pre-test, and all process variables outperform demographic variables. This provides further support for the relative lack of importance of demographic characteristics as contributors to these models of standardized test performance.

Table 12: Representative Example of Importance of Predictor Variables

Variable	MDG
Problems per workspace	883.841
Workspaces mastered per hour	815.439
Assistance per problem	737.553
Pre-test	713.693
Number of KCs mastered	633.783
Total problem solving time	571.983
Workspaces encountered	530.488
Student grade	142.592
Ethnicity	138.362
ESE status	136.895
FRPL status	133.130
LEP status	124.331

Note: As measured by mean decrease in the Gini index of impurity (MDG) in decreasing order (i.e., in order of importance) in M5 RFC model learning on 2014-15 data.

5.3. Hierarchical Models

Modelling 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, returning to the goal of regression modelling, we consider explicitly modelling this and the extent to which such considerations may improve statistical models 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 2014–2015 data set were not present in data from 2013–2014, so we use data from 2014–2015 to infer linear mixed effects models that can account for this school hierarchy. We then test these models on data from 2013–2014 and 2015–2016 and compare results to those obtained from SLR to see if explicitly modelling this hierarchy provides for significant boosts in accuracy.

5.3.1. 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,

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{1i} + \beta_2 X_{2i} + \cdots + \beta_s X_{si} + \varepsilon_i,$$
$$\beta_{0j} = \gamma_0 + \mu_{0j},$$
$$\beta_{1j} = \gamma_1 + \mu_{1j} \text{ (Eq. 8)},$$

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, $X_2, \ldots, X_s$ have fixed effects only. This model has distribution assumptions, $\beta_{0j} \sim N(\gamma_0, \tau_{00}), \beta_{1j} \sim N(\gamma_1, \tau_{11})$, and $\varepsilon_i \sim N(0, \sigma^2)$. 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 scores (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 differences between the multilevel linear models and best SLR models are the random effects.

5.3.2. Results

Table 13 provides a comparison of model performance between the best SLR model trained on 2014–2015 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 multilevel or hierarchical (i.e., linear mixed effects) models M3 through M6 outperform SLR models on the 2013-2014 test data and on models M1 through M4 on the 2015-2016 test data, but SLR models outperform the hierarchical models on M5 and M6 on 2015-2016 test data, 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 potential 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 modelling.

Nevertheless, the extent to which linear mixed effects models do outperform SLR indicates that such modeling efforts might benefit from this hierarchical approach (or investigations into how to capture these school-level effects with as-yet-unconsidered process variables).

Table 13: Comparison of the Accuracy of the Best Models Learned by Stepwise Linear Regression (SLR) and Linear Mixed Effects Modelling (LME)

	13-14		15-16	
Model	SLR	LME	SLR	LME
M1	11.837	12.912	12.830	12.647
M2	12.601	13.478	13.955	13.795
M3	12.462	11.992	13.633	13.461
M4	11.780	11.317	12.562	12.488
M5	10.893	10.347	11.436	11.507
M6	10.749	10.461	11.472	11.558

Note: Trained on data from 2014–2015 and tested on data from 2013–2014 (13–14) and 2015–2016 (15–16) using RMSE; cf. Table 6.

6. Discussion and Future Work

Our results show the potential for using formative assessment, fully integrated into learning, as a replacement for end-of-year standardized tests, but the ability to predict summative assessments based on MATHia usage has other advantages. In particular, the formative information used in these models is available to teachers immediately and diagnoses knowledge gaps and misconceptions at a much finer grain size than a standard assessment. Such data can be used to guide instructional focus and remediation. With respect to pacing, Carnegie Learning now uses a metric called the Adaptive Personalized

Learning Score (APLSE), based on these predictive models, to help teachers, students, and administrators understand whether students are on track to pass the end-of-course exam. Carnegie Learning has been able to show that students who meet the "on track" criterion on this measure are over 95% likely to receive a passing score on the final exam. Another potential use of embedded formative assessment within MATHia would be to provide a kind of Bayesian prior to the administration of a more traditional summative assessment. This might be a transitional stage, perhaps allowing a shorter summative exam than is currently possible.

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 2013–2014 training data, achieves RMSE of 11.327 on 2015–2016 test data, out-performing both SLR using 2014–2015 training data and the linear mixed effects model built on 2014–2015 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 2013–2014 data as a test set when trained on 2014–2015 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 of test scores we could expect in the best case by considering split-half and other forms of reliability of the underlying standardized test (Feng et al., 2006).

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 9 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, particularly as compared to percent correct on a traditional exam. 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. Recovering 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. Currently, high-stakes assessments are a major barrier to implementation of personalized learning, due to the administrative requirement of administering an exam on the school district's schedule, rather than the student's schedule.

Perhaps the greatest impact of this vision of assessment may be on perceptions of assessment and learning more generally. The message conveyed by the common distinction between instruction and assessment is that the assessment environment is somehow a better window into the student's knowledge than the instructional environment could be. This is almost certainly wrong. The instructional environment certainly provides a greater volume of data about student knowledge and performance than assessments. Instructional tasks are also typically more varied and rich than the kinds of tasks provided within most standardized assessments. Embedding assessment within instruction eliminates any concerns about alignment between the curriculum and the exam. Finally, the anxiety invoked by high-stakes assessments may also result in an underestimate of student ability (Lehman, Herbert, Jackson, & Grace, 2017).

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

Declaration of Conflicting Interest

Author Zheng has consulted for Carnegie Learning, Inc., and authors Fancsali, Ritter, and Berman are employed by Carnegie Learning, Inc. Carnegie Learning, Inc., is the developer and provider of MATHia software and related mathematics curricula.

Funding

This work was financially supported by an Ignite Challenge Middle and High School Math Award from the New Schools Venture Fund and by Carnegie Learning, Inc., the employer of authors Fancsali, Ritter, and Berman.

Acknowledgments

The NewSchools Venture Fund supported this work via an Ignite Challenge Middle and High School Math Award to Carnegie Learning, Inc. Carnegie Learning thanks Miami-Dade County Public Schools and Ilia Molina for providing substantial amounts of data over the past 3+ years.

References

Anozie, N. O., & Junker, B. W. (2006). Predicting end-of-year accountability assessment scores from monthly student records in an online tutoring system. AAAI Workshop on Educational Data Mining (AAAI-06), 17 July 2006, Boston, MA, USA. https://www.aaai.org/Papers/Workshops/2006/WS-06-05/WS06-05-001.pdf

Ayers, E., & Junker, B. W. (2008). IRT modeling of tutor performance to predict end-of-year exam scores. *Educational and Psychological Measurement, 68*(6), 972–987. http://dx.doi.org/ 10.1177/0013164408318758

Baker, R. S .J. d., Corbett, A. T., Roll, I., & Koedinger, K. R. (2008). Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction, 18*, 287–314. http://dx.doi.org/10.1007/s11257-007-9045-6

Baker, R. S. J. d., Gowda, S. M., Wixon, M., Kalka, J., Wagner, A. Z., Salvi, A., Aleven, V., Kusbit, G. W., Ocumpaugh, J., & Rossi, L. (2012). Towards sensor-free affect detection in Cognitive Tutor Algebra. In K. Yacef, O. Zaiane, A. Hershkovitz, M. Yudelson, & J. Stamper (Eds.), *Proceedings of the 5th International Conference on Educational Data Mining* (EDM2012), 19–21 June 2012, Chania, Greece (pp. 126–133). International Educational Data Mining Society.

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

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

Black, P. J., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice, 5*(1), 7–73. http://dx.doi.org/10.1080/0969595980050102

Bloom, B. S. (1968). Learning for mastery. *Evaluation Comment, 1*(2). https://eric.ed.gov/?id=ED053419

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140. http://dx.doi.org/10.1023/A:1018054314350

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. http://dx.doi.org/10.1023/A:1010933404324

Brookhart, S. M. (2009). Editorial: Special issue on the validity of formative and interim assessment. *Educational Measurement: Issues and Practice, 28*(3), 1–4.

Buckingham, B. R. (1921). Intelligence and its measurement: A symposium XIV. *Journal of Educational Psychology, 12*, 271–275. http://dx.doi.org/10.1037/h0066019

Campione, J. C., & Brown, A. L. (1985). *Dynamic assessment: One approach and some initial data*. Technical Report No. 361. Champaign, IL: University of Illinois at Urbana-Champaign, Center for the Study of Reading.

Carpenter, S. K., Cepeda, N. J., Rohrer, D., Kang, S. H. K., & Pashler, H. (2012). Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction. *Educational Psychology Review, 24*, 369–378. http://dx.doi.org/10.1007/s10648-012-9205-z

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction, 4*(4), 253–278. http://dx.doi.org/10.1007/BF01099821

Danks, D., & London, A. J. (2017). Algorithmic bias in autonomous systems. *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (IJCAI'17), 19–25 August 2017, Melbourne, Australia (pp. 4691–4697). Palo Alto, CA: AAAI Press.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research, 7*, 1–30.

Evans, C., & Lyons, S. (2017). Comparability in balanced assessment systems for state accountability. *Educational Measurement Issues and Practice, 36*(3), 24–34. http://dx.doi.org/10.1111/emip.12152

Feng, M., Heffernan, N. T., & Koedinger, K. R. (2006). Predicting state test scores better with intelligent tutoring systems:

Developing metrics to measure assistance required. In M. Ikeda, K. Ashlay, & T.-W. Chan (eds.), Intelligent Tutoring Systems, ITS 2006. Lecture Notes in Computer Science, vol 4053. Springer, Berlin, Heidelberg

Florida Department of Education. (2014). *FCAT 2.0 and Florida EOC Assessments Achievement Levels*. http://www.fldoe.org/core/fileparse.php/3/urlt/achlevel.pdf

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

Friedman, M. (1940). A comparison of alternative tests of significance for the problem of m rankings. *Annals of Mathematical Statistics, 11*(1), 86–92.

García, S., Fernández, A., Luengo, J., & Herrera, F. (2010). Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences*, *180*, 2044–2064. https://dx.doi.org/10.1016/j.ins.2009.12.010

Gardner, J., & Brooks, C. (2018). Evaluating predictive models of student success: Closing the methodological gap. *Journal of Learning Analytics*, *5*(2), 105–125. https://dx.doi.org/10.18608/jla.2018.52.7

Gettinger, M., & White, M. A. (1980). Evaluating curriculum fit with class ability. *Journal of Educational Psychology, 72*, 338–344. http://dx.doi.org/10.1037/0022-0663.72.3.338

Grigorenko, E. L., & Sternberg, R. J. (1998). Dynamic testing. *Psychological Bulletin*, *124*(1), 75. http://dx.doi.org/10.1037/0033-2909.124.1.75

Harlen, W., & James, M. (1997). Assessment and learning: Differences and relationships between formative and summative assessment. *Assessment in Education: Principles, Policy & Practice*, *4*(3), 365–379, http://dx.doi.org/10.1080/0969594970040304

Hart, R., Casserly, M., Uzzell, R., Palacios, M., Corcoran, A., & Spurgeon, A. (2015). *Student testing in America's great city schools: An inventory and preliminary analysis.* Washington, DC: Council of Great City Schools.

Heritage, M. (2010). *Formative assessment and next-generation assessment systems: Are we losing an opportunity?* Washington, D.C.: Council of Chief State School Officers.

Hodges, J. L., & Lehmann, E. L. (1962). Ranks methods for combination of independent experiments in analysis of variance. *Annals of Mathematical Statistics*, *33*, 482–497.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55–67.

Joshi, A., Fancsali, S. E., Ritter, S., Nixon, T., & Berman, S. (2014). Generalizing and extending a predictive model for standardized test scores based on cognitive tutor interactions. In J. Stamper et al. (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining* (EDM2014), 4–7 July 2014, London, UK (pp. 369–370). International Educational Data Mining Society. http://educationaldatamining.org/EDM2014/uploads/procs2014/posters/45_EDM-2014-Poster.pdf

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

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The knowledge–learning–instruction framework: Bridging the science–practice chasm to enhance robust student learning. *Cognitive Science*, *36*(5), 757–798. https://dx.doi.org/10.1111/j.1551-6709.2012.01245.x

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, *47*, 583–621.

Lazarín, M. (2014, October). Testing overload in America's schools. Washington, DC: Center for American Progress. https://cdn.americanprogress.org/wp-content/uploads/2014/10/LazarinOvertestingReport.pdf

Lehman, B., Hebert, D., Jackson, G. T., & Grace, L. (2017). Affect and experience: Case studies in games and test-taking. *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (CHI EA '17), 6–11 May 2017, Denver, Colorado, USA (pp. 917–924). New York: ACM. http://doi.org/10.1145/3027063.3053341

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3–62. http://dx.doi.org/10.1207/S15366359MEA0101_02

Moses, S. (2017, March 28). State testing starts today; Opt out CNY leader says changes are "smoke and mirrors." https://www.syracuse.com/schools/index.ssf/2017/03/opt-out_movement_ny_teacher_union_supports_parents_right_to_refuse_state_tests.html

Nelson, H. (2013, July). Testing more, teaching less: What America's obsession with student testing costs in money and lost instructional time. Washington, D.C.: American Federation of Teachers. http://www.aft.org/sites/default/files/news/testingmore2013.pdf

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York: Broadway Books.

Pane, J. F., Griffin, B. A., McCaffrey, D. F., & Karam, R. (2014). Effectiveness of cognitive tutor algebra I at scale. *Educational Evaluation and Policy Analysis, 36*(2), 127–144. https://dx.doi.org/10.3102/0162373713507480

Pardos, Z. A., Baker, R. S., San Pedro, M., Gowda, S. M., & Gowda, S. M. (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. https://doi.org/10.18608/jla.2014.11.6

Pardos, Z. A., Heffernan, N. T., Anderson, B., Heffernan, C. L., & Schools, W. P. (2010). Using fine-grained skill models to fit student performance with Bayesian networks. In C. Romero, S. Ventura, M. Pechenizkiy, & R. S. J. d. Baker (Eds.), *Handbook of educational data mining* (pp. 417–426). Boca Raton, FL: CRC Press.

Pashler, H., Cepeda, N. J., Wixted, J., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 3–8. http://doi.org/10.1037/0278-7393.31.1.3

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).

Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice*, *28*(3), 5–13. https://dx.doi.org/10.1111/j.1745-3992.2009.00149.x

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163.

Razzaq, L., Feng, M., Nuzzo-Jones, G., Heffernan, N. T., Koedinger, K. R., Junker, B., ... & Livak, T. (2005). Blending assessment and assisting. In C.-K. Looi, G. I. McCalla, B. Bredeweg, & J. Breuker (Eds.), *Proceedings of the 12th International Conference on Artificial Intelligence in Education* (AIED 2005), 18–22 July 2005, Amsterdam, The Netherlands (pp. 555–562). Amsterdam, The Netherlands: IOS Press.

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*, 181–210. https://dx.doi.org/10.1111/j.1745-6916.2006.00012.x

Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive Tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review, 14*(2), 249–255. https://dx.doi.org/10.3758/BF03194060

Ritter, S., Joshi, A., Fancsali, S. E., & Nixon, T. (2013). Predicting standardized test scores from Cognitive Tutor interactions. In S. K. D'Mello et al. (Eds.), *Proceedings of the 6th International Conference on Educational Data Mining* (EDM2013), 6–9 July 2013, Memphis, TN, USA (pp. 169–176). International Educational Data Mining Society/Springer.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464. http://dx.doi.org/10.1214/aos/1176344136

Shute, V. J., & Ke, F. (2012). Games, learning, and assessment. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in game-based learning*. New York: Springer. http://dx.doi.org/10.1007/978-1-4614-3546-4

Shute, V. J., Levy, R., Baker, R., Zapata, D., & Beck, J. (2009). Assessment and learning in intelligent educational systems: A peek into the future. In S. D. Craig & D. Dicheva (Eds.), *Proceedings of the 14th International Conference on Artificial Intelligence in Education* (AIED '09), *Vol. 3: Intelligent Educational Games*, 6–10 July 2009, Brighton, UK (pp. 99–108). Amsterdam, The Netherlands: IOS Press.

Shute, V. J., & Moore, G. R. (2017). Consistency and validity in game-based stealth assessment. In H. Jiao & R. W. Lissitz (Eds.), *Technology enhanced innovative assessment: Development, modeling, and scoring from an interdisciplinary perspective* (pp. 31–51). Charlotte, NC: Information Age Publishing.

Sirin, S. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*(3), 417–453. http://www.jstor.org/stable/3515987

Slade, S., & Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. *American Behavioral Scientist*, *57*(10), 1509–1528. https://dx.doi.org/10.1177/0002764213479366

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 263–331). New York: American Council on Education, Macmillan.

State of Minnesota. (2017). *Standardized student testing: 2017 evaluation report*. Office of the Legislative Auditor. http://www.auditor.leg.state.mn.us/ped/pedrep/studenttesting.pdf

Tagami, T. (2018, February 15). Alternative testing bill passes Georgia Senate. *Politically Georgia.* https://politics.myajc.com/news/state--regional-education/alternative-testing-bill-passes-georgia-senate/AANXBjII7whHivfY2QuCKI/

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B, 58*(1), 267–288. https://dx.doi.org/10.1111/j.2517-6161.1996.tb02080.x

US Department of Education. (2016). Table 215.30: Enrollment, poverty, and federal funds for the 120 largest school districts, by enrollment size in 2014 (selected years, 2013–14 through 2016). *Digest of Education Statistics*. National Center for Education Statistics. https://nces.ed.gov/programs/digest/d16/tables/dt16_215.30.asp

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York: Springer.

Walker, T. (2018, January 4). Educators strike big blow to overuse of standardized testing in 2017. *neaToday: News and Features from the National Education Association*. http://neatoday.org/2018/01/04/educators-strike-big-blow-to-the-overuse-of-standardized-testing/

Wiliam, D. (2011). *Embedded formative assessment*. Bloomington, IN: Solution Tree Press.

Zerr, C. L., Berg, J. J., Nelson, S. M., Fishell, A. K., Savalia, N. K., & McDermott, K. B. (2018). Learning efficiency: Identifying individual differences in learning rate and retention in healthy adults. *Psychological Science*, *29*(9). http://dx.doi.org/10.1177/0956797618772540