

# Predicting Standardized Test Scores from Cognitive Tutor Interactions

Steve Ritter, Ambarish Joshi, Stephen E. Fancsali, Tristan Nixon

Carnegie Learning, Inc.

437 Grant Street, Suite 918

Pittsburgh, PA 15219

1-412-690-2442

{sritter, ajoshi, sfancsali, tnixon}@carnegielearning.com

## ABSTRACT

Cognitive Tutors are primarily developed as instructional systems, with the goal of helping students learn. However, the systems are inherently also data collection and assessment systems. In this paper, we analyze data from over 3,000 students in a school district using Carnegie Learning's Middle School Mathematics tutors and model performance on standardized tests. Combining a standardized pretest score with interaction data from Cognitive Tutor predicts outcomes of standardized tests better than the pretest alone. In addition, a model built using only 7th grade data and a single standardized test outcome (Virginia's SOL) generalizes to additional grade levels (6 and 8) and standardized test outcomes (NWEA's MAP).

## Keywords

Cognitive Tutors, Assessment, Mathematics.

## 1. INTRODUCTION

Cognitive Tutors are primarily developed as instructional systems, with a focus on improving student learning. While the systems continually assess student knowledge with respect to a set of underlying knowledge components [6], the standard for effectiveness of an educational system is usually taken to be the ability of that system to produce improved performance on an external measure, typically a standardized test.

Carnegie Learning's Cognitive Tutors for mathematics have done well on such measures [13, 15, 18] but, in these studies, the tutoring system has been, essentially, treated as a black box. We know that, as a whole, students using a curriculum involving the Cognitive Tutor outperformed students using a different form of instruction on standardized tests, but we don't know what specific aspects of tutor use were associated with improved performance. An understanding of the process variables (time, errors, hint usage and other factors) that are correlated with learning can provide us with insight into the specific student activities that seem to lead to learning. Another perspective on the Tutor is that, if we are able to strongly correlate Cognitive Tutor data with standardized test data, then the Cognitive Tutor itself may be considered an assessment, which is validated with respect to the external standardized test. In addition, to the extent that we can identify process variables that predict external test scores, we can provide guidance to teachers as to expectations for their students on the state examinations.

In most cases, Carnegie Learning does not have access to student-level outcome data on standardized tests. For the study reported here, we partnered with a school district in Eastern Virginia. The district provided Carnegie Learning with student data for all 3224 middle school students who used Cognitive Tutor in the district

during the 2011/12 school year. The data included student demographics and outcomes on the Virginia Standards of Learning (SOL) assessment and NWEA's Measures of Academic Progress (MAP) assessment. The district also provided MAP scores from the Fall of 2011, which provides information about student abilities prior to their encounter with the Cognitive Tutor. This dataset allows us to explore which particular behaviors within Cognitive Tutor are associated with improved outcomes and to test whether we are better able to predict outcomes knowing student behaviors within the tutor than we could be able to predict from demographic and prior knowledge variables.

While other analyses [5, 14] have modeled outcomes based on tutor process data, the current analysis goes beyond previous efforts by building a model based on a single grade level and outcome and then applying the model to two outcome measures across three grade levels.

## 1.1 Virginia's Standards of Learning (SOL) Assessment

The Virginia Department of Education's Standards of Learning (SOL) provide minimum expectations for student knowledge for several subjects at the end of each grade level or after specific courses [21]. Students take standardized tests based on the mathematics SOL annually for grades 3 through 8 as well as after taking particular courses after grade 8 (e.g., after taking Algebra I). The SOL exam includes multiple choice items as well as "technology enhanced" items that may include drag-and-drop, fill-in-the-blank, graphing and "hot spot" identification in a picture.

With the advent of the No Child Left Behind Act, there is great interest in developing predictive models of student performance on high-stakes tests like the SOL mathematics assessment to identify students that may need remediation. Cox [8], for example, develops regression models to predict grade 5 SOL mathematics assessment scores using student scores on Mathematics Curriculum Based Measurement (M-CBM) benchmark assessments, which are shorter, formative assessments that track student progress over time. This study reports that M-CBM alone can account for roughly 40% to 60% of the variance in SOL assessment scores across three Virginia school districts.

## 1.2 NWEA's Measures of Academic Progress® (MAP) Assessment and RIT score

NWEA's MAP is a computer-based adaptive assessment. MAP assessments deliver student scores on the Rasch Unit (RIT) scale, an equal interval measurement, intended to provide scores comparable across grade levels and to help track student progress from year to year [20]. For the district in our study, the MAP assessment was administered in both the fall and spring semesters. NWEA recently published data correlating MAP performance

with Virginia SOL assessment performance for over 5,000 students. For 2012 data they report correlations (as Pearson's  $r$ ) of 0.759, 0.797, and 0.75 between MAP scores and SOL mathematics assessment scores for grades 6, 7, and 8, respectively [19]. These figures are comparable to the correlations we report here between RIT and SOL in grades 6 and 7, but the correlation for grade 8 was lower ( $r$  of 0.755, 0.704 and 0.551 for grades 6, 7, 8, respectively).

### 1.3 Cognitive Tutor

The Cognitive Tutor presents a software curriculum as a sequence of "units," which are major topics of instruction. Units are divided into one or more sections, which represent subtopics. Each section is associated with one or more skills (or knowledge components), which are the target of the mastery learning. Each section has a large pool of available problems (from 10s to 1000s, depending on the section), which are chosen to remediate students on each of the target skills. The Tutor considers a student to have mastered a section when Bayesian Knowledge Tracing [7] judges that there is at least a 95% probability that the student knows the skill. When a student masters all of the skills in a section, the system allows the student to proceed to the next section (or unit, if this is the final section in a unit). When the student completes a problem without mastering all of the skills for that section, the Tutor picks a new problem for the student, focusing on the skills that still need to be mastered. Students require different numbers of problems to master all of their skills.

Although the intent of the system is for students to progress through topics as they master them, we recognize that there will always be some students who are not learning from the tutor's instruction (or who are learning too slowly). For this reason, each section specifies a maximum number of problems. If the student reaches the maximum without mastering all skills in the section, the student is advanced to the next section without mastery. Teachers are notified of this advancement in reports. Thus, in our models, we make a distinction between sections (or skills) encountered and sections (or skills) mastered. Teachers may also manually move a student to another point in the curriculum, which would also result in an encounter with a section (or skill) without mastery.

Problems within the Cognitive Tutor are completed in steps, each of which represents a discrete entry into the system (such as filling in a text field). The tutor evaluates each step and provides immediate feedback. Students can also ask for help on each step. The number of steps required to complete a problem depends on the complexity of the problem and the particular strategy that the student uses to complete the problem. Some problems may be completed in 5-10 steps but others may require 30 or more. Within a section, problems typically require the same (or very similar) number of steps. Because the variability of problems across sections affects the potential for making errors and asking for hints (as well as the expected time to complete the problem), we normalize hints, errors and time within each section and then calculate an average across sections, for each student. This normalization also helps account for the fact that different students encountered different sections (depending on grade level and custom sequences of units), so we should have different expectations about time, hints and errors per problem within the sections they did encounter.

The software includes various instructional activities and supports in addition to the Cognitive Tutor. These include basic lesson text, multiple-choice tests, step-by-step examples and other components. None of these other activities directly affect our

assessment of mastery. For this reason, we distinguish between "problem time" (the time that students spend within Cognitive Tutor problems) and the total time that students spend logged in to the tutor.

## 2. DISTRICT IMPLEMENTATION

The students in this study used Cognitive Tutor software developed for middle school mathematics (grades 6, 7 and 8) in 12 schools in the district. The software was used by 3224 students: 1060 in sixth grade; 1354 in seventh grade and 810 in eighth grade.

Carnegie Learning delivers software sequences aligned to educational standards for these grades. The sixth grade sequence contains 45 units and 131 sections; the seventh grade sequence contains 34 units and 92 sections and the eighth grade sequence contains 37 units and 88 sections. The school district also created software sequences targeted towards students who were performing below grade level, as part of a Response to Intervention (RTI) implementation [10].

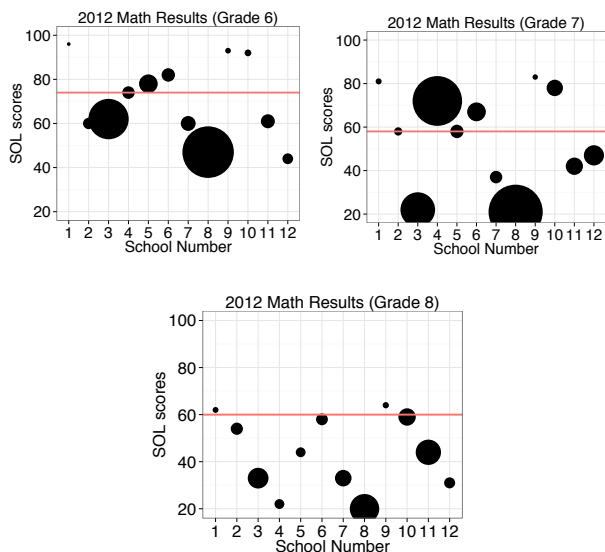
Carnegie Learning recommends that, when used as a basal school curriculum, students use the Cognitive Tutor as part of their math class two days/week. Assuming 45-minute classes and a 180-day school year, this would result in approximately 54 hours of use. Due to scheduling issues, absenteeism and other factors, it is common for districts to average only 25 hours, however. RTI implementations may involve more intensive (5 days/week) practice on prerequisite skills, typically completed in an RTI class which takes place in parallel with work in the basal class. Thus, students in an RTI sequence may be asked to work on the Tutor twice as much (or more) than those who are not in an RTI sequence. On the other hand, if students in the RTI class are able to demonstrate mastery of the target material, they are removed from that class, so the RTI class does not necessarily last all year.

The Tutor is available to students through a browser, so they can go home (or elsewhere) and continue work that they began at school. However, many students do not use the Tutor outside of class and so, for most students, the amount of time that they spend with the tutor is dictated by the frequency with which their teacher tells them to use the software.

Our analysis does not distinguish between students who used the Tutor in an RTI capacity, as a basal curriculum or in some other capacity.

Across the schools and grade levels in our data set, usage varied widely, from median of 1.05 hours in grade 8 in one school to a median of 29.86 hours in grade 7 in a different school.

Figure 1 provides a schematic for understanding overall performance of students at the schools involved in the study. There are three graphs, representing the three grade levels. Each school is represented by a dot, with the size of the dot proportional to the number of students in that school and grade level. The vertical position of the dot represents the school's overall 2012 SOL score. The horizontal line represents the state average for the grade level. The figure shows that students in the district reflect a range of abilities relative to the state, with students somewhat underperforming the state mean in all grades.



**Figure 1: School-level SOL math results for the schools in the study. The size of the dot represents the number of students in the study. The horizontal line represents the state average.**

### 3. ANALYSIS APPROACH

Our primary purpose in this work is to build a general model that can be used to predict outcomes from standardized tests based on usage and behaviors within Cognitive Tutor. To this end, we build a model based on a subset of the students and a single outcome variable and then test the model on the rest of the students and an additional outcome variable. Since the seventh grade cohort was the largest in our data, we chose to build the model on the seventh graders. We chose SOL as the outcome measure in building our model, because it is the most important outcome to the schools using Cognitive Tutor.

Since we expected use of Cognitive Tutor to influence outcomes only if there was some substantial use of Cognitive Tutor over the school year, we excluded students who, in total, spent less than 5 hours using Cognitive Tutor. This reduced the number of students considered by the model from 1354 7<sup>th</sup> graders (and 3224 students in all grades) to 940 7<sup>th</sup> graders (and 2018 students overall).

In order to explore the influence of different kinds of information, we constructed 5 models:

- M1 – includes only RIT pretest score as a predictor
- M2 – includes only Cognitive Tutor process variables
- M3 – includes the Cognitive Tutor process variables used in M2, plus student demographics
- M4 – includes RIT pretest score plus student demographics
- M5 – includes M3 variables, plus RIT pretest score

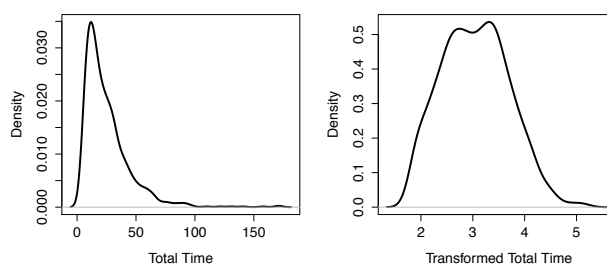
To build M2, we used stepwise regression to identify the Cognitive Tutor process variables, of those considered (see Variable Selection, below), that minimized the Bayesian Information Criterion (BIC). This model considered only 7<sup>th</sup> grade data and only SOL as an outcome. To build M3, we included those variables selected for M2 and student demographic variables and used stepwise regression (using BIC) to choose the most predictive demographic variables, keeping the process variables

fixed. We used 10-fold cross-validation of these models to ensure that we were not overfitting, but the cross-validated models were very similar to the model found by simple stepwise regression, so we only consider the stepwise regression model here.

### 3.1 Cognitive Tutor Process Variables

Since overall usage of the Tutor varied substantially between students, we reasoned that a good model would take into account variables that represent this overall usage (either as represented by time or by various completion measures). Since most usage of the tutor is controlled by the teacher (or the class that the student is in), variance within a class might be better captured by metrics representing activities taken within a problem.

Many of the variables we considered were highly skewed in the data, and so, following common practice, we applied log transforms to them. For example, Figure 2 (left) shows the distribution of time spent on the tutor (excluding students who spent fewer than 5 hours). Although the median of the distribution is 19.8 hours, there is a long tail of students who spend a much longer time using the tutor. A log transform produces the more normal distribution shown in Figure 2b.



**Figure 2: Distribution of total time on the tutor (left graph) and the log-transformed distribution (right graph).**

Since we are modeling students that completed different mathematics topics (different units and sections within Cognitive Tutor), we normalized many of these process variables within section. This normalization results in variables that represent the difference between a student’s behavior and the average student’s behavior, in standard deviation units, so that values are comparable across sections of the curriculum. In cases where we log transformed the data, normalization followed the log transformation. To compute a value for each student, we averaged the normalized (z) scores across all of the sections that the student encountered. For “aggregate” variables, which sum across sections of the curriculum, we normalized with respect to the distribution of data across the full school year. Since we normalize all variables in the model, the magnitude of the coefficients gives a sense of the relative impact of each variable.

Based on some preliminary correlations with outcomes in these data and other datasets, we considered 13 Cognitive Tutor process variables for our model:

#### Aggregate variables

- **Total\_time**: the total amount of time that the student was logged in to the tutor. This variable was log transformed.
- **Total\_problem\_time**: the amount of time that students spent on the problem-solving activities within the tutor. This differs from total\_time by excluding time spent reading lesson content, viewing worked examples and

several other activities. This variable was log transformed.

- **Percent\_non\_problem\_time**: the percentage of time that the student was logged in to the software but did things other than solving problems in Cognitive Tutor. This variable was log transformed.
- **Sections\_encountered**: the total number of sections attempted by the student. This variable was log transformed.
- **Skills\_encountered**: the total number of skills within the sections that the student encountered. This variable was log transformed.
- **Percent\_skills\_mastered**: the percentage of skills encountered that reached mastery. This variable was subtracted from 1 and log transformed.
- **Percent\_sections\_mastered**: the percentage of sections encountered that were mastered. This variable was subtracted from 1 and log transformed.
- **Sections\_mastered\_per\_hour**: This is the average number of sections mastered by the student for each hour of time spent on problems. We consider this variable to be an indicator of the efficiency with which the student uses the system. Efficiency may be affected by the student's prior knowledge and conscientiousness but also by the extent to which teachers are effective in assisting students when they get stuck. This variable was log transformed.

#### Section-normalized variables

- **Time\_per\_problem**: the average amount of time spent on each problem. This variable was log transformed and normalized
- **Hints\_per\_problem**: the average of the number of hints requested in each problem. This variable was log transformed and normalized.
- **Errors\_per\_problem**: the average of the number of errors committed in each problem. This variable was log transformed and normalized.
- **Assistance\_per\_problem**: the average of the sum of the number of hints and the number of errors in each problem. This variable is an indicator of the extent to which students struggle to complete problems and was log transformed and normalized.
- **Problems\_per\_section**: the number of problems required to master the skills in the section (or, for non-mastered sections, the maximum number of problems in the section). This variable was log transformed and normalized.

### 3.2 Demographic Variables

In addition to process variables, we considered the following student demographic variables (note that statistics are calculated based on the 2018 students with more than 5 hours usage):

- **Sex**: male or female. In our sample, there were 1027 boys and 991 girls.
- **Age**: in days (as of 06/01/12). In 6<sup>th</sup> grade,, the mean was 4493 with standard deviation 163. In 7<sup>th</sup> grade, mean was 4868 with standard deviation of 172. In 8<sup>th</sup>

grade, the mean was 5269, with standard deviation of 184.

- **Lunch\_status**: this is a common proxy for socio-economic status. We coded this as a binary variable indicating whether students were eligible for free or reduced lunch prices. 72.5% of students were in this category.
- **Limited\_English\_Proficiency**: A binary variable coding whether the student was identified as having a poor mastery of English. 6.9% of students in our sample were identified as being in this category.
- **Race**: The school provided coding in six categories, shown here:

Description	Students
American Indian	20
Asian	72
Black/African American	1064
White	785
Hawaiian /Pac. Islander	5
Multi-racial	72

- **Hispanic\_origin**: A binary variable representing whether the student is of Hispanic origin. 17.1% of students were identified as Hispanic.
- **Special\_education\_status**: We coded this status as representing four categories: no special status, learning disability, physical disability or other (which included students with multiple disabilities and those that were listed as special ed but not classified). 9.6% of students were identified with a learning disability, 8.4% with physical disability and 1% with other.

## 4. RESULTS

### 4.1 Fitted Models

The process variables found in M2 and included in M3 and M4 were *total\_problem\_time*, *skills\_encountered*, *sections\_encountered*, *assistance\_per\_problem* and *sections\_mastered\_per\_hour*. A summary of the standardized model (M2) coefficients is shown in Table 1.

**Table 1: Cognitive Tutor process variables and standardized coefficients included in the models predicting SOL.**

Variable	Coefficient	p value
assistance_per_problem	-0.351	<2e-16
sections_encountered	0.422	0.004028
sections_mastered_per_hour	0.390	6.71E-10
skills_encountered	-0.456	0.000141
total_problem_time	0.258	0.000502

Three variables (*total\_problem\_time*, *skills\_encountered* and *sections\_encountered*) represent aggregate usage of the tutor. *Skills\_encountered* is entered with a negative coefficient, perhaps trading off with the (positive) *sections\_encountered* and indicating that students benefitted from completing a larger number of sections, particularly if the sections involved a

relatively small number of skills. Although it is tempting to interpret sections with small numbers of skills as simpler (or shorter) sections, the number of skills tracked in a section is not always a measure of complexity, particularly in sections that include a number of skills that many students have previously mastered [17].

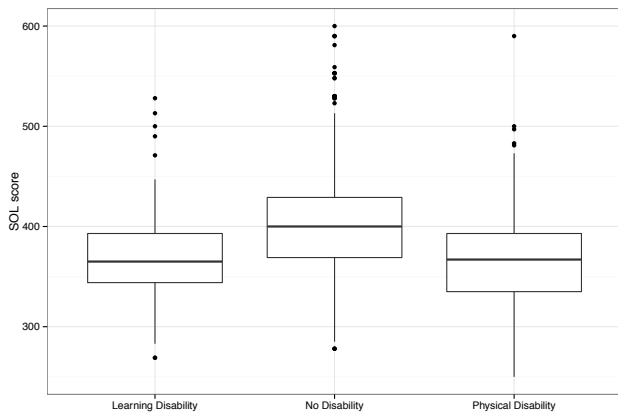
The model also includes *assistance\_per\_problem*, and *sections\_mastered\_per\_hour*. Together, these variables reflect students' ability to efficiently work through problems, staying on task, making few errors and not overly relying on hints.

Demographic variables found in M3 and used in M4 and M5 included *lunch\_status* and *Age*. Student of low socio-economic status (i.e., those that qualify for free and reduced-price lunches indicated by *lunch\_status*) perform significantly worse on the SOL, even after accounting for Cognitive Tutor process variables. *Age* is also a factor in the model indicating that, after accounting for other model variables, students who are older tend to underperform their younger classmates. Although significant, the age effect is small. One year increase in age, within grade, results in a score reduction of 8 points. Consider that 8-point difference relative to the larger differences in school-level SOL scores shown in Figure 1.

**Table 2: Variables and standardized coefficients for M3 applied to SOL**

Variable	Coefficients	p value
assistance_per_problem	-0.340	<2e-16
sections_encountered	0.369	0.011183
sections_mastered_per_hour	0.368	4.04E-09
skills_encountered	-0.403	0.000663
total_problem_time	0.240	0.001043
lunch_status	-0.106	3.30E-05
age	-0.071	0.003737

We were surprised to find that special education status was not a significant predictor. Figure 3 shows why: in these data, student SOL scores do not vary much with respect to special education status.



**Figure 3: SOL score by special education status.**

Table 3 shows a summary of the complete M5. Once RIT is included, *age* and *sections\_encountered* are no longer significant predictors, although all other predictors are still significant.

A summary of the model fits is shown in Table 4. It is notable that *RIT\_pretest* predicts SOL scores somewhat better than the Cognitive Tutor process model ( $R^2$  of 0.50 for M1 vs. 0.43 for M2) and that adding demographics to either Cognitive Tutor process variables (M3) or RIT alone (M4) increases the predictive validity of the model only slightly. The combination of RIT and Cognitive Tutor process variables (M5) increases the fit of the model substantially, compared to either M3 or M4. This may indicate that the RIT pretest and the Cognitive Tutor process variables are capturing different and complementary aspects of student knowledge, motivation and preparedness.

Despite containing the most variables, M5 is the only model that shows a substantially lower BIC score than M1 (RIT alone), indicating that these variables, in combination, provide substantial explanatory power.

**Table 3: Variables and standardized coefficients for M5, as applied to SOL**

Variable	Coefficients	p value
assistance_per_problem	-0.134	1.02E-05
sections_encountered	0.188	0.141868
sections_mastered_per_hour	0.272	7.80E-07
skills_encountered	-0.262	0.011882
total_problem_time	0.219	0.000669
lunch_status	-0.070	0.00166
age	-0.028	0.197271
RIT pretest	0.476	<2e-16

**Table 4: Summary of fits for models of SOL**

Model	Number of variables	BIC	$R^2$
M1 (RIT)	1	2041.451	0.50
M2 (CT)	5	2181.015	0.43
M3 (CT+Demog)	7	2167.764	0.45
M4 (RIT+Demog)	3	2030.582	0.51
M5 (Full)	8	1928.369	0.57

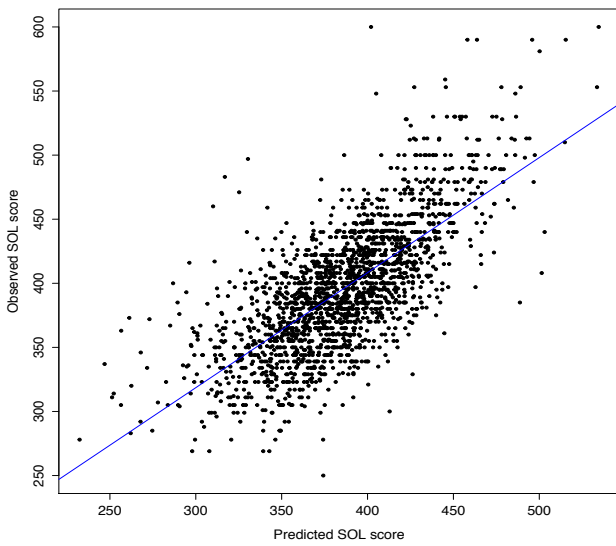
## 4.2 Generalizing to other grades

Table 5 shows fit for the models as applied to sixth and eighth grade students' SOL scores (and as applied to the full population of students), using the variables and coefficients found by fitting the seventh grade student data. The model fits the sixth grade data remarkably well, with an  $R^2$  of 0.62, higher even than the fit to the seventh grade data. The fit to eighth grade data is not as strong, with an  $R^2$  of 0.32. This may be due to both the smaller original population in eighth grade and the relatively low usage. Median usage for eighth graders was only 16.3 hours (as opposed to 20.3 hours in sixth grade), and only 438 students (54%) of eighth graders used the tutor for more than five hours.

**Table 5: Summary of fits of the model as applied to SOL for held-out students (grades 6 and 8) and to the whole population (grades 6,7,8)**

R <sup>2</sup>	Grade 6	Grade 8	All grades
M1	0.57	0.30	0.40
M2	0.46	0.18	0.38
M3	0.46	0.18	0.42
M4	0.57	0.30	0.43
M5	0.62	0.32	0.51

Figure 4 demonstrates the fit of M5 to the SOL data for all grade levels.



**Figure 4: Relationship of predicted SOL using M5 to actual SOL scores for students at all grade levels.**

Although the fit is very good ( $R^2=0.51$ ), the model appears to be slightly underpredicting SOL scores for the best students and slightly overpredicting SOL scores for the worst students.

### 4.3 Generalizing to the RIT posttest

Our next question was whether the variables we found for modeling SOL would also produce a good model of RIT score. In order to do this, we used the variables from M5 and regressed the model against the RIT posttest score, again fitting the seventh grade students. Table 6 shows the resulting standardized coefficients.

Note that, with RIT as the outcome variable, *sections\_mastered\_per\_hour*, *total\_problem\_time* and *age* are no longer significant predictors for the model. It may be that RIT, as an adaptive test, imposes less time pressure on students, since there is no apparent set of questions to be completed in a fixed amount of time.

**Table 6: Standardized coefficients and p values for M5, predicting RIT posttest scores**

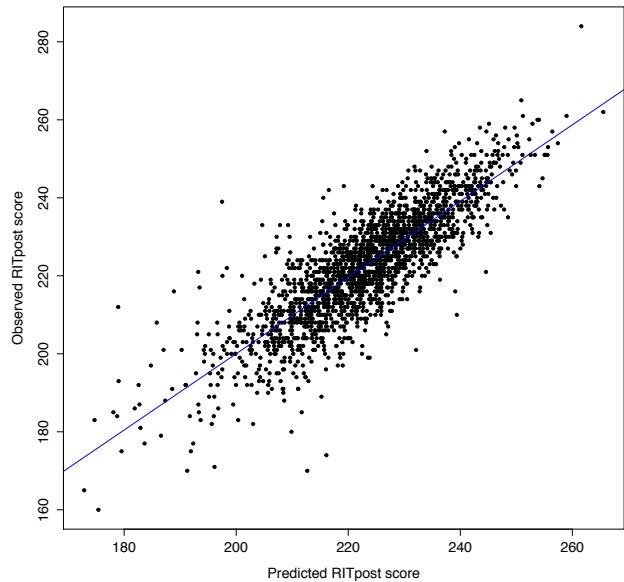
Variable	Coefficients	p value
assistance_per_problem	-0.186	1.68E-15
sections_encountered	0.267	0.006
sections_mastered_per_hour	0.031	0.45747
skills_encountered	-0.206	0.00927
total_problem_time	0.003	0.95771
lunch_status	-0.044	0.0092
age	0.008	0.64494
RIT pretest	0.677	<2e-16

Table 7 shows fits for the five models, as applied to the RIT posttest. As expected, RIT pretest predicts RIT posttest better than the RIT pretest predicts the SOL posttest ( $R^2$  for M1/SOL for 7<sup>th</sup> grade is 0.50 vs. 0.72 for M1/RIT). Even given this good fit for M1, process variables and demographics significantly improve the model, reducing BIC from 1502 to 1409.

**Table 7: R<sup>2</sup> and BIC for models as applied to RIT posttest.**

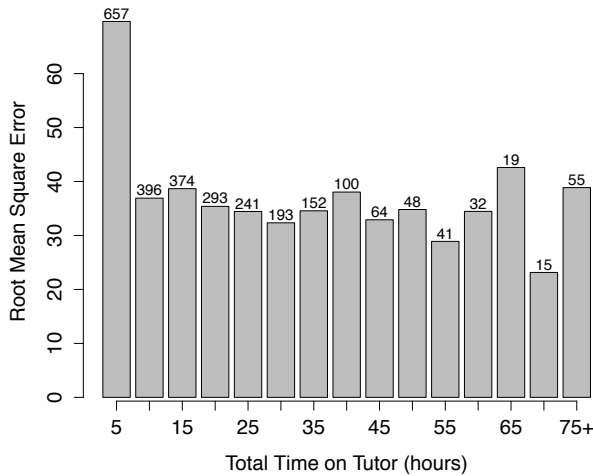
	R <sup>2</sup>				BIC
	Grade 6	Grade 7	Grade 8	All	Grade 7
M1	0.67	0.72	0.59	0.68	1502.128
M2	0.46	0.49	0.26	0.41	2085.297
M3	0.48	0.50	0.27	0.40	2077.530
M4	0.68	0.72	0.59	0.68	1501.578
M5	0.71	0.75	0.60	0.71	1408.899

The fit to the full population is illustrated in Figure 5.



**Figure 5: Relationship of predicted RIT posttest scores using M5 to actual RIT posttest scores for students at all grade levels.**

To this point, we have been considering the full population to be the population of students who used Cognitive Tutor for at least 5 hours. Figure 6 shows how the model applies to all students who used the Cognitive Tutor for any period of time.



**Figure 6: M5 Root Mean Square Error (from predicted SOL outcome), as a function of total time on the Tutor. Numbers above the bars represent the number of students represented by the bar.**

Figure 6 demonstrates that the model’s ability to predict SOL scores decreases dramatically for students with low Cognitive Tutor usage. This is to be expected: outcomes for students who used the software very infrequently are only lightly influenced by anything they may have learned from the Tutor, and the relatively small amount of data that the Tutor was able to collect from such students represents a noisy estimate of the student’s abilities. Our choice of 5 hours as a cutoff point was based on experience with other data sets. Figure 6 validates this cutoff for our data.

## 5. DISCUSSION

The work presented here provides a good model of how we might use Cognitive Tutor, either with or without additional data, to predict student test outcomes on standardized tests. The model was able to generalize to different student populations, and the variables found for a model to predict SOL provided strong predictions of RIT as well.

Surprisingly to us, demographic factors proved to be relatively unimportant to our models.

Since we were able to improve on the RIT pretest model by adding Cognitive Tutor process variables, our efforts show that such variables provide predictive power beyond that provided by a standardized pretest, even when the pre- and post-test are identical (as in the case with the RIT outcome). A consideration of the types of information that may be contained in Cognitive Tutor data but not in pretest data provide us with guidance on how we might extend this work and improve our model. We will consider 5 broad categories of factor: learning, content, question format, process and motivation.

*Learning:* The most obvious difference between the RIT pretest and the Cognitive Tutor process variables is that the RIT provide information about students at the beginning of the school year, while Cognitive Tutor data is collected throughout the year. One extension of this work in exploring the role of learning would be

to look at how well the model presented here would predict outcomes if we only considered data from the first six months (or three months – or less) of the school year. If such an early prediction model were to work, it could act as an early warning system for teachers and administrators [1].

*Content:* Although both RIT and Cognitive Tutor are designed to align to the SOL, it is possible that differences in that alignment are responsible for some of the improvement that Cognitive Tutor provides over RIT alone in predicting SOL scores. Feng et al.[9], using the ASSISTment system, built a one-parameter IRT model to predict test outcomes, an approach which allows them to weight different ASSISTment items better with respect to the outcome variable. Our models considered all skills and sections to have equivalent predictive power, but a more sophisticated model could take content alignment into account.

*Question format and problem solving:* Cognitive Tutor problems involve multiple steps and the kind of strategic decision making that is characteristic of problem solving. Traditional standardized tests are multiple choice or single-step fill-in-the-blank and tend to assess more procedural and less conceptual knowledge. In past research [13], Cognitive Tutor students have shown stronger performance (relative to control) on open-ended responses than on traditional standardized tests. Part of the prediction within SOL may be due to the closer alignment between Cognitive Tutor and the technology-enhanced SOL question types. Most states in the United States (but not Virginia) have adopted the Common Core State Standards. Two consortia, Partnership for Assessment of Readiness for College and Careers (PARCC) and Smarter Balanced Assessment Consortium (SBAC), are developing new assessments to align with these standards. Both consortia are including substantial non-traditional items, including some multi-step problem solving. In order to align with the Common Core assessments, the 2012-13 version of MAP includes technology-enhanced items, similar to those in SOL. It remains to be seen whether such a change would account for some of the variance now explained by Cognitive Tutor variables in our models.

*Process:* By process, we mean the way that students go about working in Cognitive Tutor. In the SOL models, the strongest Cognitive Tutor predictor of outcomes was the number of sections completed per hour. We characterize this variable as coding some kind of efficiency; it captures students who are able to get to work and stay on task (and also who are able to generally succeed at the task). Unlike most standardized tests, online systems like Cognitive Tutor have the ability to take both time and correct performance into account.

In models of both SOL and RIT, a strong predictor was the amount of assistance (hints and errors) required by the student. Although assistance can be an indicator of lack of knowledge (leading to large numbers of hints and errors), it may also be an indicator of lack of confidence (in students who ask for hints rather than risk making an error) or gaming the system.

In this paper, we have only considered variables aggregated at the level of problem. For example, our data considers the number of hints and errors per problem but not the pattern or timing of those hints and errors. This simplification was required because more detailed data were not available for all students in this data set. However, other work [e.g. 2, 3] has shown that more detailed “detectors” of gaming and off-task behavior, which rely on patterns and timing of actions within Cognitive Tutor, can be strong predictors of outcomes. We would expect such detectors to be more sensitive to student behavior than the relatively coarse-grained measures used here.

*Motivation and non-cognitive factors:* Much recent work has pointed to the powerful effect that attitudes towards learning can have on standardized test outcomes [11, 12]. Pardos et al. [16] were able to use detectors of student affect (including boredom, concentration, confusion, frustration) to predict standardized test outcomes. Such affect detectors have already been developed for Cognitive Tutor [4] and, in principle, could be added to our models.

While we are very encouraged with the results that we have seen in this paper, we recognize that more detailed data may provide us better ability to predict student test outcomes from Cognitive Tutor data.

## 6. ACKNOWLEDGMENTS

This work was supported, in part, by LearnLab (National Science Foundation), award number SBE-0836012. Michael Yudelson provided helpful suggestions on our approach to analysis.

## 7. REFERENCES

- [1] Arnold, K.E. 2010. Signals: Applying Academic Analytics. *Educause Quarterly*, 33, 1.
- [2] Baker, R.S.J.d. 2007. Modeling and Understanding Students' Off-Task Behavior in Intelligent Tutoring Systems. In *Proceedings of ACM CHI 2007: Computer-Human Interaction*, 1059-1068.
- [3] Baker, R.S., Corbett, A.T., Koedinger, K.R. 2004. Detecting Student Misuse of Intelligent Tutoring Systems. *Proceedings of the 7th International Conference on Intelligent Tutoring Systems*, 531-540.
- [4] Baker, R.S.J.d., Gowda, S.M., Wixon, M., Kalka, J., Wagner, A.Z., Salvi, A., Alevan, V., Kusbit, G., Ocumpaugh, J., and Rossi, L. 2012. Towards Sensor-Free Affect Detection in Cognitive Tutor Algebra. In *Proceedings of the Fifth International Conference on Educational Data Mining*, 126-133.
- [5] Beck, J. E., Lia, P. and Mostow, J. 2004. Automatically assessing oral reading fluency in a tutor that listens. *Technology, Instruction, Cognition and Learning*, 1, pp.61-81.
- [6] Corbett, A.T. and Anderson, J. R. 1992. Student modeling and mastery learning in a computer-based programming tutor. In C. Frasson, G. Gauthier and G. McCalla (Eds.), *Intelligent Tutoring Systems: Second international conference proceedings* (pp. 413-420). New York: Springer-Verlag.
- [7] Corbett, A.T., Anderson, J.R. 1995. Knowledge tracing: modeling the acquisition of procedural knowledge. *User Modeling & User-Adapted Interaction* 4, (1995), 253-278.
- [8] Cox, P.A. 2011. *Comparisons of Selected Benchmark Testing Methodologies as Predictors of Virginia Standards of Learning Test Scores*. Doctoral Thesis. Virginia Polytechnic Institute and State University.
- [9] Feng, M., Heffernan, N.T., & Koedinger, K.R. 2009. Addressing the assessment challenge in an Online System that tutors as it assesses. *User Modeling and User-Adapted Interaction: The Journal of Personalization Research (UMUAI journal)*. 19(3), 243-266, August, 2009.
- [10] Gersten, R., Beckmann, S., Clarke, B., Foegen, A., Marsh, L., Star, J. R., & Witzel, B. 2009. *Assisting students struggling with mathematics: Response to Intervention (RTI) for elementary and middle schools* (NCEE 2009-4060). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from <http://ies.ed.gov/ncee/wwc/publications/practiceguides/>
- [11] Good, C., Aronson, J., & Harder, J. A. 2008. Problems in the pipeline: Stereotype threat and women's achievement in high-level math courses. *Journal of Applied Developmental Psychology*, 29, 17-28.
- [12] Good, C., Aronson, J., & Inzlicht, M. 2003. Improving Adolescents' Standardized Test Performance: An Intervention to Reduce the Effects of Stereotype Threat. *Journal of Applied Developmental Psychology*, 24, 645-662.
- [13] Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. 1997. Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, 8, 30-43.
- [14] McCuaig, J. and Baldwin, J. 2012. Identifying successful learners from interaction behaviour. EDM 2012
- [15] Pane, J., Griffin, B. A., McCaffrey, D. F. and Karam, R. 2013. Effectiveness of Cognitive Tutor Algebra I at scale. RAND Working Paper WR-984-DEIES.
- [16] Pardos, Z., Baker, R.S.J.d., San Pedro, M., Gowda, S.M. & Gowda, S.M. 2013. Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes.
- [17] Ritter, S., Harris, T. H., Nixon, T., Dickison, D., Murray, R. C. and Towle, B. 2009. Reducing the knowledge tracing space. In Barnes, T., Desmarais, M., Romero, C., & Ventura, S. (Eds.) *Educational Data Mining 2009: 2nd International Conference on Educational Data Mining, Proceedings*. Cordoba, Spain.
- [18] Ritter, S., Kulikowich, J., Lei, P., McGuire, C.L. & Morgan, P. 2007. What evidence matters? A randomized field trial of Cognitive Tutor Algebra I. In T. Hirashima, U. Hoppe & S. S. Young (Eds.), *Supporting Learning Flow through Integrative Technologies* (Vol. 162, pp. 13-20). Amsterdam: IOS Press.
- [19] Northwest Evaluation Association 2012. Virginia linking study: a study of the alignment of the NWEA RIT scale with the Virginia Standards of Learning (SOL). <http://www.nwea.org/sites/www.nwea.org/files/resources/V A 2012 Linking Study.pdf>
- [20] Northwest Evaluation Association 2013. Computer-Based Adaptive Assessments. Retrieved February 23, 2013. <http://www.nwea.org/products-services/assessments/>
- [21] Virginia Department of Education 2013. Standards of Learning (SOL) & Testing. Retrieved February 23, 2013. <http://www.doe.virginia.gov/testing/>