Generalizing and Extending a Predictive Model for Standardized Test Scores Based on Cognitive Tutor Interactions

Ambarish Joshi, Stephen E. Fancsali, Steven Ritter, Tristan Nixon, Susan R. Berman
Carnegie Learning, Inc.
437 Grant Street, Suite 918
Pittsburgh, PA 15219 USA
888.851.7094
{ajoshi, sfancsali, sritter, tnixon, sberman}@carnegielearning.com

ABSTRACT

Recent work demonstrates that process data from intelligent tutoring systems (ITSs) can be used to predict student outcomes on high-stakes, standardized tests. Such models are important if ITSs are to be used for formative assessment and as replacements for external assessments. Recent work used various measures of learning efficiency and performance from problem-level, aggregate data from Carnegie Learning's Cognitive Tutor to predict standardized test scores on the state of Virginia's Standards of Learning exam. We generalize this model to a different school district, state, and standardized test and examine extending the model using finer-grained data.

Keywords

Formative assessment, standardized tests, intelligent tutoring systems, Cognitive Tutor, off-task behavior, gaming the system

1. INTRODUCTION & BACKGROUND

Advanced learning systems like Cognitive Tutor (CT) [8] and ASSISTments [4], which assess students as they teach, have the potential to reduce time taken away from instruction to assess student knowledge. By fully integrating instruction with assessment, they ensure that the two are well aligned. Recent work has aimed to demonstrate correlations between such unconventional assessments and exams to determine whether they prepare students for assessments or could replace conventional high-stakes assessments (e.g., [5,7,9]).

The CT mathematics intelligent tutoring system (ITS) is known to improve student learning and performance on standardized tests (e.g., [6]), and recent work [9] demonstrated that a model incorporating CT process data predicts middle school outcomes on Virginia's (VA's) Standards of Learning (SOL) exam [10]. We generalize this result to a new population of students from a different school district and U.S. state on a different test. We also investigate extending the model by using finer-grained data.

We ask: what process data should instructional/analytics systems track, and what level of granularity (e.g., problem-level vs. problem-solving steps) is sufficient for tracking? Both questions are important if ITSs are to be used for instruction *and* formative assessment, and if instructor dashboards and diagnostic systems are to be useful and scalable.

1.1 Cognitive Tutor

CT curricula are sequences of topical sections in instructional units. Sections contain a set of problems, each of which targets one or more knowledge components (KCs). Problems are adaptively presented to students according to KCs that a student has yet to master, as probabilistically assessed by CT. Students proceed to the next section when they master all KCs in a section. A student can choose to ask for a hint at any problem-solving step. As assessing KC mastery depends on errors made and hints requested, students may require different numbers of problems.

1.2 Previous Work & VA's SOL Exam

Past work has associated learning system process data with standardized test scores. For example, data from ASSISTments (e.g., counts of help requests) have been used to predict Massachusetts Comprehensive Assessment System (MCAS) scores (e.g., [5]). Problem-level features used in these models do not require logging data at the level of individual student actions, but they still provide satisfactory models of test scores.

Predictions of sensor-free, data-driven "detectors" of gaming the system [2], off-task behavior [1], and affect [3] with ASSISTments data have been used to predict MCAS scores [7]. Detectors require fine-grained tracking of student actions in ITSs rather than features like aggregate counts of hints. A natural question is whether finer-grained data provide information about test scores beyond that provided by problem-level data.

Previous work [9] predicted outcomes of VA's SOL test from problem-level CT process data. Data included usage for 3,224 students in Grades 6-8 across 12 schools. Grade 7 data were used to build an ordinary least squares (OLS) linear regression model. Five variables were significant: (1) *Total Problem Time* - problem-solving time; (2) (number of) *Skills Encountered* (3) (number of) *Sections Encountered*; (4) *Assistance Per Problem* - average sum of # of hints requested and errors made for each problem; and (5) (average number of) *Sections Mastered Per Hour*. Model parameter estimates for Grade 7 data (model adjusted $R^2 = 0.43$) were used to predict outcomes for Grade 6 ($R^2 = 0.46$), Grade 8 ($R^2 = 0.18$), and overall ($R^2 = 0.38$).

2. GENERALIZING THE MODEL

2.1 West Virginia's WESTEST 2

We generalize the model that predicted SOL scores by modeling data from a school district in West Virginia (WV) that uses a different standardized test, WESTEST 2. For math, WESTEST 2 assesses a student's on defined standards, objectives, and skills, using multiple-choice questions and gridded response items [11].

2.2 Data & Results

We build models of data from the 2012-2013 school year, including usage information for 636 students, mostly 9th graders taking Algebra 1, with 5+ hours of CT usage and scores above the "novice" ranking for WESTEST 2 achievement descriptors, as

students with a novice ranking are likely from a different learner sub-population than that which we target here.

Our approach starts with the variables found in the prior work. S*kills Encountered* and *Sections Encountered* are highly correlated (r = 0.989), so we disregard the variable with lower correlation to WESTEST 2 scores (*Sections Encountered*). Building a stepwise regression model from remaining variables, including neither *Skills Encountered* nor *Total Problem Time* improves the model over that of Table 1 (R^2 = 0.295). Student-level (10-fold) cross validation does not lead to models that differ substantially, so two variables generalize to WV's exam.

Table 1: Standardized regression coefficients, & significance for generalized model of WESTEST 2 scores (*p < .001)**

Variable	Coefficient
Assistance Per Problem	-0.225***
Sections Mastered Per Hour	0.372***

3. EXTENDING THE MODEL

Data-driven "detectors" of gaming the system [2] (e.g., abusing hints or excessive guessing), off-task behavior [1] and affect [3] have been used to predict MCAS scores [7]. Detectors use features "distilled" from problem-solving-step-level (i.e., finer-grained data) logs.

We construct *Steps Gamed* and *Steps Offtask* variables using data for 5 million+ student actions, to capture the proportion of student problem-solving steps detected as instances of these behaviors. Table 2 reports the regression model (R^2 = 0.322) including these variables and correlations to WESTEST 2 outcomes.

Table 2: Regression coefficients for extended model & correlations with learning outcome (p <.01; ***p<.001)**

Variable	Regression Coefficient	Correlation with WESTEST 2
Assistance Per Problem	-0.07	-0.47***
Sections Mastered Per Hour	0.396***	0.52***
Steps Gamed	-0.224***	-0.51***
Steps Offtask	0.129**	-0.2***

Measures of student efficiency, gaming the system, and off-task behavior are significant predictors of outcomes. Off-task behavior is relatively weakly correlated with WESTEST 2 outcomes. *Assistance Per Problem* and *Steps Gamed* are highly correlated (r = 0.8, two-tailed p < .001); if gaming is a common cause of more assistance (i.e., hint abuse) and less learning, conditional on *Steps Gamed*, *Assistance Per Problem* and learning would be independent, so *Assistance Per Problem* would be insignificant. *Sections Mastered Per Hour* is negatively correlated with *Steps Gamed* (r = -0.68, p < .001) and *Steps Offtask* (r = -0.59, p < .001), so gaming and off-task behavior seem to provide the same information about outcomes as variables in the generalized model.

4. DISCUSSION

Two features from a model that predicts VA's SOL test generalize to WV's WESTEST 2. We attempted to extend the model by including features that require finer-grained data about problem-solving steps; we find that gaming the system and off-task behavior do not substantially improve our predictions for WESTEST 2, in part because assistance is highly correlated with gaming. Our results thus suggest that the benefits of collecting fine-grained data needed to construct sophisticated features may not be substantial for use in test score prediction. Nevertheless, engineered features and fine-grained data may provide for real-time assessment to target interventions.

5. ACKNOWLEDGMENTS

Sujith M. Gowda and Ryan S.J.d. Baker provided both code for and assistance with detector models used in this work.

6. REFERENCES

[1] 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* (San Jose, CA, April 28 – May 3, 2007). ACM, New York, 1059-1068.

[2] Baker, R.S.J.d., de Carvalho, A. M. J. A. 2008. Labeling student behavior faster and more precisely with text replays. In *Proceedings of the 1st International Conference on Educational Data Mining* (Montreal, 2008). 38-47.

[3] 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 *Proceedings of the 5th International Conference on Educational Data Mining* (Chania, Greece, 2012). 126-133.

[4] Feng, M., Heffernan, N.T., Koedinger, K.R. 2009. Addressing the assessment challenge in an intelligent tutoring system that tutors as it assesses. *User Model. User-Adap.* 19 (2009), 243-266.

[5] 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 *Proceedings of the 8th International Conference on Intelligent Tutoring Systems* (Jhongli, Taiwan, 2006). 31-40.

[6] Pane, J., Griffin, B. A., McCaffrey, D. F., Karam, R. 2014. Effectiveness of Cognitive Tutor Algebra I at scale. *Educ Eval Policy An* 36 (2014), 127-144.

[7] Pardos, Z.A., Baker, R.S.J.d., San Pedro, M.O.C.Z., Gowda, S.M., and Gowda, S.M. 2013. Affective states and state tests: Investigating how affect throughout the school year predicts end of year learning outcomes. In *Proceedings of the Third International Learning Analytics and Knowledge Conference* (Leuven, Belgium, 2013). ACM, New York, NY, 117-124. DOI= http://doi.acm.org/10.1145/2460296.2460320

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

[9] Ritter, S., Joshi, A., Fancsali, S.E., Nixon, T. 2013. Predicting standardized test scores from Cognitive Tutor interactions. In *Proceedings of the 6th International Conference on Educational Data Mining* (Memphis, TN, 2013). 169-176.

[10] Virginia Department of Education. 2014. Standards of Learning (SOL) & Testing. Retrieved February 23, 2014. http://www.doe.virginia.gov/testing/

[11] West Virginia Department of Education, Office of Assessment and Accountability. 2014. WESTEST 2 Overview. Retrieved February 23, 2014. http://wvde.state.wv.us/oaa/westest_index.html