Predicting end-of-year accountability assessment scores from monthly student records in an online tutoring system

Nathaniel O. Anozie and Brian W. Junker
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
{noa,brian}@stat.cmu.edu

Abstract

We present methodology for developing functions that predict student scores on end of year state accountability exams from dynamic testing metrics developed from intelligent tutoring system log data. Our results confirm the findings of Heffernan et al. (2006) that online tutoring log based metrics provide better predictions than using paper and pencil benchmark tests. Our approach provides a family of prediction functions to be used throughout the year, in order to provide timely and valid feedback to teachers and schools about student progress. Since the same dynamic testing metrics are used in all prediction functions in the family, we can also begin to understand the changing influence of these metrics on prediction over time.

Introduction

Recently there has been intense interest in using periodic benchmark tests to predict student performance on end-of-year accountability assessments (Olson 2005). Benchmark tests are typically paper-and-pencil tests given at regular intervals, from three times a year to monthly, in order to predict progress toward proficiency on state accountability exams. Some benchmark tests also try to function as formative assessments for teachers, so that the classroom time occupied by the benchmark test is not completely lost to teachers' instructional mission. Nevertheless, teachers may still find the time needed for benchmark tests to be an intrusion on instructional time.

An alternative approach may be available when an online, computer-based tutoring system is in place. The benefits of online tutoring systems are well known: (Koedinger et al. 2000) study classroom evaluations of the Cognitive Tutor Algebra course (e.g., Koedinger et al. 1997; Koedinger et al. 2000) and demonstrate that students in tutor classes outperform students in control classes by 50–100% on targeted real-world problem-solving skills and by 10–25% on standardized tests. (Beck et al. 2004) argue for a mixed predictive/formative use for benchmarks based on tutor interaction logs for their online reading tutor. In this paper we will explore Beck's suggested approach with a different tutoring system.

Heffernan et al. 2001 have developed an online tutoring system for eigth grade mathematics that is explicitly aligned to state exam standards, and in fact takes released state exam questions and "morphs" of them as the main student tasks for tutoring. Their system is called the *ASSISTment* system, since it implicitly *assesses* progress toward proficiency on the state exam at the same time it *assists* student learning in mathematics.

In this paper we design effective linear prediction models for end of year performance from tutoring system data using cross-validation based variable selection. In order to provide predictions regularly throughout the year, we build separate prediction models for each month of the school year, based on current and previous months' summaries of student activity. In order to preserve comparability and interpretability across months, we constrain the model building so that the same variables are used in each prediction model.

After this introduction we describe our model building approach in more detail, and then we describe in detail the data we used in this work. Following this we present the results of our analyses and consider the changing influence of dynamic testing metrics on prediction over time. Finally we explore the possibilities of using our models for online test score prediction, as well as possible extensions to our methodology.

Our work builds on and greatly extends analyses of (Lee and Ting 2005), based on an earlier 6–8 week pilot of the ASSISTment system. It is also closely related to the work of our colleagues Feng, Heffernan and Koedinger (in press; 2006) to find a single regression model to predict state exam scores from a full-year summary of ASSISTment data. Recently (Pardos et al. 2006) and (Feng et al. 2006) have examined per-skill and per-item performance metrics (rather than the aggregated metrics we consider here) to help refine prediction of end-of-year performance. Our work is aimed at cases where fine-grained per-skill and per-task metrics may not be readily available; it also allows an examination of the changing influence of dynamic assessment metrics over time on end-of-year performance predictions.

Modeling

ASSISTment dynamic testing metrics

Eighth graders from two schools in the Worcester Public School District used the ASSISTment system in the September thru July 2004-2005 school year. Four hundred "main questions" or main items were available for students to practice with. These main items were mostly previous Massachusetts Comprehensive Assessment System (MCAS) [1] exam questions and *prima facie* equivalent versions (or "morphs") of them. Each main item was supplemented with a set of "scaffold questions" intended to help students who did not get the main item right on the first try. An ASSISTment is one main item together with its available "scaffold questions". In October students spent on average 27 minutes in the system and answered approximately 11 main questions and 21 scaffold questions.

Thus not only can the system keep track of right and wrong answers on first attempts at MCAS questions, it can also differentiate between students who need more or less help to get the right answers: students who ask for more hints, who take longer before answering a question, etc., may be distinguished from students who ask for few hints or need only a brief time to answer. These dynamic testing metrics (cf. Campione et al. 1985) can be very predictive of student achievement. The dynamic testing metrics we considered were constructed by examining the earlier work of (Lee and Ting 2005) and Feng, Heffernan and Koedinger (2006; in press), and by considering additional variables that may be helpful in predicting MCAS scores (see Table 1).

In our approach we first construct monthly summaries of these variables, for each month of the ASSISTment field trial of 2004–2005. For each month, we build a linear model predicting MCAS scores from that month's summaries as well as any previous month's summaries. Variable selection for the models proceeds in two stages, first eliminating variables that are highly redundant with other variables in Table 1, and then selecting a final set of variables by minimizing cross-validation prediction error (Wasserman 2004, pp. 218, 364). The final set of monthly prediction models are constrained to use the same set of variables, in order to facilitate comparison of the models across months.

Model building

Stage 1 In Stage 1 our goal is to eliminate variables that are highly correlated with other variables in Table 1. Separately for each month, we begin with the full set of monthly summaries and calculate variance inflation factors (VIF; see Hamilton 1992, pp. 133, 135) for each variable. The variable with the highest VIF is eliminated, and VIF's are recalculated. This backwards-selection procedure is iterated until all VIF's are less than 10 (this is equivalent to multiple-$R^2 < 0.90$ for predicting each remaining variable from the others). For a variable to be considered in Stage 2, it must be retained in all trial months.

[1] http://www.doe.mass.edu/mcas

Stage 2 In Stage 2 we peform cross-validation-assisted forward selection where we evaluate prediction error by 10-fold cross validation (Wasserman 2004, pp. 220, 363; see also Snee 1977), using the sum over all models of the mean absolute deviation (MAD). This forward selection procedure is not like standard forward selection in two ways. First, we are evaluating variables for inclusion in seven prediction models simultaneously: each variable is either included in all models or it is excluded from all models. Second, each variable actually appears seven times, summarizing the same dynamic testing metric in each of the seven trial months, and all current and past monthly summaries are considered in each monthly prediction model.

Data: the 2004-2005 ASSISTment tutor trial

Of the 912 students that worked with the ASSISTment system at some time during the trial year, only 105 students had complete data in each of the seven months October through April.

We imputed students' monthly summaries for months in which they did not use the ASSISTment system by copying forward their summaries from the most recent month in which they did use the system. Hence students considered must have worked with the ASSISTment system in September and/or October 2004. We think this imputation is a reasonable reflection of student performance because it was rare to go back more than two months to retrieve data for imputation: in March where the most number of imputations were made, 868 of the 912 students used the ASSISTment system. 50% of these March students needed no imputation, 23% needed pulling forward from February, and 14% needed pulling forward from January, etc.

After imputation 697 students of the total 912 students had complete data. 560 of these 697 students had usable data for the variable selection procedure we described above (they had MCAS scores and never had a completion time of zero seconds on any main question or scaffold that they answered). 15 of these 560 students had perfect main question scores and so required additional, logical, imputation for percent correct on scaffolds (100%), time spent on incorrectly answered scaffolds (zero seconds), etc. 362 of these 560 students had complete pre- and post-tests, as described by Feng et al. (2006), and some of our analysis focus on this subset.

Results

Stage 1.

After running the Stage 1 backwards elimination procedure, 11 variables remained in the pool. These variables are listed in bold in Table 1. The choices made by this procedure generally make sense; for example among variables measuring quantity of questions completed, rate of question completion, and time spent on questions, at least one of these was eliminated.

Nevertheless two variables that we felt might have an important interpretive contribution were also eliminated from the variable pool: Percent correct on main questions (PctCorMain), and Number of hints requested plus number of in-

Table 1: Variables in bold face passed our Stage 1 collinearity check, and were considered for Stage 2 variable selection. Italicized variables were also considered in Stage 2, because of their strong substantive interpretation; they did not substantially increase collinearity when added to the variable pool for Stage 2.

Summary Per Month	Definition	Average Value October (Most Used Month)
$PctCorMain$	Percent of main questions correct	0.28
PctCorScaf	percent of scaffolds correct	0.41
SecIncScaf	Number of seconds on incorrect scaffolds	784.20
NumPmAllScaf	Number of complete scaffolds per minute	1.36
NumHintsIncMainPerMain	Hints plus incorrect main questions per ASSISTment	3.46
AttIncMain	Number of attempts on incorrect main questions.	8.63
SecCorScaf	Number of seconds on correct scaffolds	181.00
NumPmAllMain	Number of complete main questions per minute	1.13
PctSecIncScaf	Percent of time on scaffolds spent on incorrect scaffolds	0.79
SecIncMain	Number of seconds on incorrect main questions.	466.00
MedSecIncMain	Median number of seconds per incorrect main question	65.11
PctSecIncMain	percent of time on main questions spent on incorrect main questions	0.78
$NumHintsIncMain$	Hints plus incorrect main questions	32.54
NumAllMain	Number of complete main questions	11.26
NumAllScaff	Number of complete scaffolds	20.77
NumCorMain	Number of correct main questions	4.26
NumHintsAll	Number of hints on main questions and scaffolds	25.54
NumAttAll	Number of attempts	48.59
NumSecAll	Number of seconds on main questions and scaffolds	1613.00
AttCorMain	Number of attempts on correct main questions	4.25
AttCorScaf	Number of attempts on correct scaffolds	8.23
AttIncScaf	Number of attempts on incorrect scaffolds	27.48
SecCorMain	Number of seconds on correct main questions	181.40
NumCorScaf	Number of correct scaffolds	8.42
MedSecAllMain	Median number of seconds per main question	56.90
NumIncMain	Number of incorrect main questions	7.00
NumIncScaf	Number of incorrect scaffolds	12.35

correct main questions (NumHintsIncMain) These two variables were added back into the variable pool and VIF's were recomputed. They are listed in italics in Table 1. After their re-introduction, none of the eleven previously selected variables had a max VIF larger than 10. Thus, all 13 bold-faced and italicized variables in Table 1 were available for analysis in Stage 2.

Stage 2.

Our main analysis was conducted on the full set of 560 students described in the data and subjects section, and the 13 bold-faced and italicized variables listed in Table 1. Variables were added one-at-a-time, according to maximum reduction in cross-validation MAD summed over all seven linear prediction models; when no further variable additions reduced the MAD, the procedure ended.

Figure 1 shows this cross-validation based variable selection procedure, averaged over 100 cross-validation runs. On each run random splits of the data were made. Variables from Table 1 are listed across the bottom of the graph; the number in parentheses following each variable is the number of runs (out of 100) for which the variable was selected into the regression models. The solid line graphed in Figure 1 shows the average order of selection of each variable, across the runs for which it was selected: an average order of one indicates that the given ASSISTment metric, on average, was selected first into all seven models; two indicates that the given ASSISTment metric, on average, was selected second into all seven. Approximate Wald-style 95% confidence intervals for order of selection are also shown.

Figure 1 shows that percent correct variables: PctCorMain and Percent correct on scaffold questions (PctCorScaf), and a time efficiency variable: number of seconds spent on incorrectly answered scaffolds (SecIncScaf) appeared as the first, second and third variable entered in each of the 100 cross-validation runs. A second time efficiency variable number of scaffolds completed per minute (NumPmAllScaff) and a help seeking variable number of hints and incorrect main questions per number of main questions were entered fourth and fifth in about 78 of the 100 cross-validation runs.

In Figure 2 we show the results of a second cross-validation experiment designed to explore how predictions of the MCAS exam improve as more data is accumulated. In this experiment, variables were added to each of the seven regression models one at a time, in the order indicated in Figure 1. 10-fold cross-validation MAD's were calculated after each variable was added. This procedure was also repeated 100 times, and the resulting MAD's were averaged.

In Figure 2, the top most line represents the October model. The first point on this line represents the average 10-fold cross-validation MAD, averaged over 100 cross-validation replications, for a model containing only the October summary of PctCorMain. The second point on the same line represents the average MAD, for a model containing only October summaries PctCorMain and PctCorScaf,

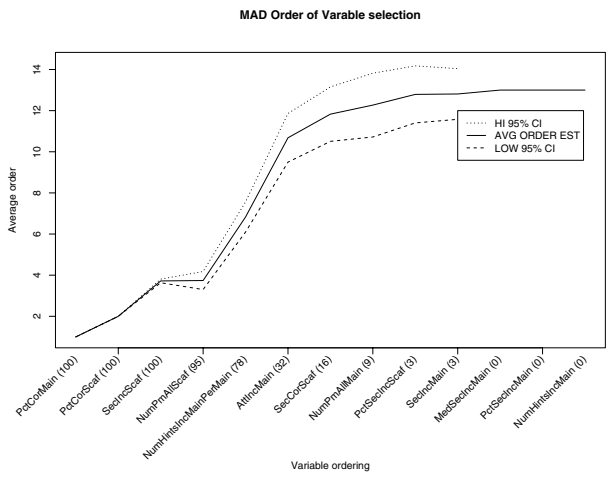

Figure 1: Average order of variable selection in stage 2. The number in parenthesis indicates the number of trials the variable was selected. Each average order score is calculated by averaging the order of inclusion in each of the indicated number of trials. An order of 1 indicates the variable is entered first on average, a 2 indicates if the variable is entered second on average.

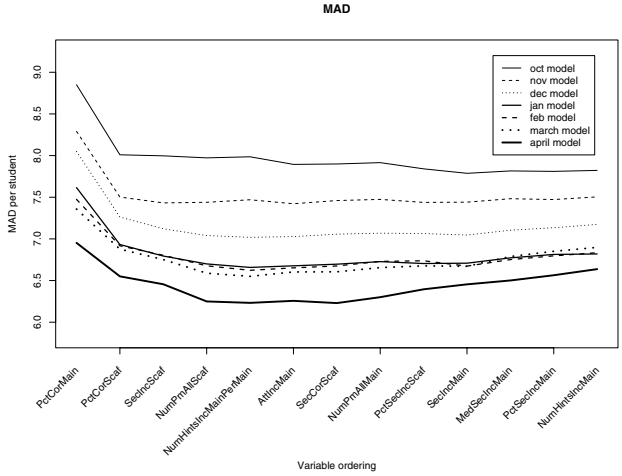

Figure 2: Average MAD score produced per model. Each point is calculated by averaging over 100 crossvalidation experiments. The order of variable entry is indicated on the horizontal axis and was determined from the first 100 run cross-validation experiment. See Figure 1.

and so forth: as more variables are added, the MAD goes down. On the other hand, the first point on the second line in Figure 2 represents the average MAD for a model containing October and November summaries of the PctCorMain variable etc.

For our final prediction models we included all variables until the last variable for which the April model's average MAD score decreased. Thus, we consider all variables added before and including this cutoff, they are: percents correct on main items and on scaffold items, rates of completion for scaffold items, time spent on incorrectly answered scaffold items and number of hints plus incorrect main questions per main question.

Comparison with Bench mark tests

Here we compare prediction of MCAS scores using the variables chosen in our variable selection procedure, with prediction using only the paper and pencil pre- and post-tests described in (Feng et al. 2006), for the subset of 362 of the full set of 560 students who also have these paper and pencil test scores. For this analysis we did not use cross validation to calculate mean absolute deviations; instead we calculated training-sample MAD's by producing model fits determined by the subset of 362 students and appropriate ASSISTment metrics or paper test variables. Residuals without cross validation should be lower than residuals using cross validation, because with cross validation we are not using each students data in each model fit.

Table 2 shows that predictions of the MCAS exam get better as we accumulate data, that is, we see that R^2 adjusted increases and MAD's decrease with additional ASSISTment metric monthly summaries. In addition we see that three months of ASSISTment data (October through December) produce models with better R^2 adjusted and MAD than the pretest only model; and four months' data exceeded the predictive power of the pre-test and post-test together.

Coefficients

Table 3 gives coefficients of dynamic testing metrics for the seven linear models. We see how the influence of various online system metrics on MCAS score prediction change over time. As would be expected, percent correct on main questions and scaffold questions contribute positively to predict MCAS score, each time their coefficients are significantly different from zero in these models. In addition, number

Table 2: R^2 adjusted, and (MAD) for 362 students who have pre and post paper tests. Stage 2 models using top five variables are shown below Pre-test and Post-test. These variables are: PctCorMain, PctCorScaf, SecIncScaf, NumPmAllScaf, and NumHintsIncMainPerMain.

Model	R^2 adjusted	MAD	% Error (MAD/54)*100
Pretest	0.523	6.690	12.388
Pre-test & Post-test	0.557	6.440	11.925
October	0.368	7.737	14.328
Oct thru November	0.474	7.069	13.090
Oct thru December	0.527	6.585	12.195
Oct thru January	0.568	6.151	11.392
Oct thru February	0.578	6.013	11.135
Oct thru march	0.577	5.978	11.070
Oct thru April	0.637	5.462	10.115

Table 3: Coefficients of dynamic testing metrics for the seven linear models. Coefficients significantly different from zero at the $\alpha = 0.05$ level are in bold.

Model	PctCorMain		PctCorScaf		SecIncScaf *1000		NumPmAllScaf		NumHintsInc -MainPerMain		Intercept	
	Coef	Se	Coef	Se	Coef	Se	Coef	Se	Coef	Se	Coef	Se
Oct	**14.66**	**2.63**	**19.37**	**2.59**	-1.22	0.74	**1.34**	**0.54**	-0.13	0.28	14.01	14.66
Oct	**5.58**	**2.73**	**11.92**	**2.70**	**-2.07**	**0.72**	0.63	0.55	-0.03	0.28	**12.35**	**5.58**
Nov	**11.32**	**2.68**	**12.05**	**2.45**	**-1.04**	**0.52**	0.55	0.42	0.03	0.28	–	–
Oct	2.44	2.69	**8.67**	**2.65**	**-2.09**	**0.70**	0.52	0.54	-0.07	0.27	4.71	2.44
Nov	**6.33**	**2.72**	**7.89**	**2.50**	**-1.66**	**0.58**	0.17	0.44	-0.18	0.30	–	–
Dec	**11.47**	**2.66**	**12.36**	**2.37**	1.09	0.61	**1.41**	**0.55**	**0.88**	**0.34**	–	–
Oct	1.87	2.56	**6.50**	**2.54**	**-1.77**	**0.67**	0.47	0.52	-0.09	0.26	0.44	1.87
Nov	**5.23**	**2.60**	**7.39**	**2.38**	**-1.63**	**0.56**	0.08	0.42	-0.19	0.29	–	–
Dec	4.51	2.71	**6.71**	**2.47**	1.00	0.60	**1.33**	**0.55**	0.44	0.35	–	–
Jan	**16.03**	**2.54**	**8.81**	**2.12**	0.01	0.59	0.22	0.44	**0.99**	**0.35**	–	–
Oct	2.22	2.57	**6.14**	**2.54**	**-1.69**	**0.67**	0.50	0.52	-0.06	0.26	-1.92	2.22
Nov	4.66	2.60	**7.09**	**2.38**	**-1.40**	**0.58**	0.19	0.43	-0.21	0.29	–	–
Dec	4.54	2.69	**6.09**	**2.46**	1.05	0.60	**1.24**	**0.55**	0.43	0.35	–	–
Jan	**13.49**	**2.69**	**8.50**	**2.27**	-0.09	0.64	-0.18	0.48	**0.89**	**0.38**	–	–
Feb	**6.41**	**2.46**	1.41	2.03	-0.59	0.81	**0.85**	**0.42**	0.30	0.32	–	–
Oct	1.88	2.56	**6.15**	**2.53**	**-1.74**	**0.66**	0.59	0.52	-0.04	0.26	-3.15	1.88
Nov	4.18	2.59	**6.03**	**2.38**	**-1.32**	**0.59**	0.31	0.43	-0.26	0.29	–	–
Dec	4.43	2.68	**6.39**	**2.47**	1.03	0.59	1.00	0.55	0.41	0.35	–	–
Jan	**12.21**	**2.72**	**7.70**	**2.30**	0.15	0.65	-0.21	0.48	**0.84**	**0.38**	–	–
Feb	3.29	2.66	-0.09	2.27	-0.57	0.88	0.58	0.50	0.21	0.36	–	–
March	**7.60**	**2.70**	2.94	2.43	-0.30	0.69	0.59	0.49	0.32	0.28	–	–
Oct	1.76	2.43	3.98	2.42	**-1.87**	**0.63**	0.62	0.49	0.06	0.25	**-4.72**	**1.76**
Nov	3.65	2.45	5.27	2.25	**-0.88**	**0.56**	0.08	0.41	-0.23	0.28	–	–
Dec	**5.33**	**2.54**	**6.72**	**2.35**	0.89	0.57	1.02	0.52	0.56	0.34	–	–
Jan	**7.31**	**2.66**	**5.71**	**2.20**	-0.02	0.63	-0.09	0.46	0.39	0.37	–	–
Feb	2.00	2.53	0.51	2.16	-0.80	0.84	0.43	0.48	0.27	0.35	–	–
March	**6.62**	**2.65**	1.18	2.40	-0.23	0.66	0.21	0.48	0.48	0.29	–	–
April	**7.83**	**1.95**	**6.32**	**1.70**	0.96	0.80	**1.41**	**0.47**	-0.15	0.19	–	–

of scaffolds completed per minute and number of hints and incorrect main questions per number of main questions contribute positively to MCAS score prediction: Thus, a higher rate of completion of scaffold questions seems to be evidence of proficiency, rather than cheating or some other gaming behavior. On the other hand, number of seconds spent on incorrectly answered scaffold questions contributes negatively in all of the cases in which its coefficients are significantly different from zero. Both the ASSISTment system, and teachers themselves, may wish to devote more attention to students who are going slowly through scaffold questions; further analysis may be needed to determine whether these students spend more time per scaffold, or encounter more scaffolds, than their peers.

Some trends in the regression coefficients can be understood from Table 3. We see that the most recent summaries of percent correct on main questions seem to be most important for predicting MCAS scores, in contrast to percent correct on scaffolds where early summaries appear to be significant. In addition we see that time spent on incorrectly answered scaffolds seems to matter more in the early months. The last two variables: number of completed scaffolds per minute and number hints and incorrect main questions per main question seem to matter in later months.

Two other aspects of Table 3 are worth noting. First, even when the coefficients are not significant, the magnitudes of the coefficient for the two percent correct scores tends to be higher for the later months' summaries in each model (this seems especially true for main questions, less so for scaffolds). Thus the linear models are to some extent "downweighting" old information about each student's performance, in making MCAS predictions. Second, and an exception to the first observation, coefficients for January summaries tend to be large and/or significant in any model in which they appear. We are not sure what makes January special in these analyses.

Conclusion

In this paper we have presented a methodology for developing functions that predict student scores on end of year state accountability exams from dynamic testing metrics developed from intelligent tutoring system log data. Although collecting data through the ASSISTment system takes longer than collecting paper and pencil benchmark testing data, our analysis agreed with that of (Feng et al. 2006) in that predictions by the ASSISTment system out performed predictions by a bench mark test. The first of these variables is a direct ASSISTment system analogue to paper and pencil test scores. The other four variables are uniquely available in the ASSISTment environment. It is especially interesting to see that efficiency in completing scaffold questions appears in all but two of the variables considered; such data is not easy to collect except in the context of a computer-based tutoring and testing system.

In this paper we begin to understand the changing influence of these metrics on prediction over time. Understanding how these metrics relate to learning as time changes is important in developing models that could potentially inform schools about state standards. For example the Proficiency Index [2] for Massachussets schools can be easily derived from these models.

We also repeated the analyses described in this paper using mean squared error (MSE) and classification error as our cross-validation criterion and we saw simlar results. For example we used the five achievement levels used by Massachussets to group student performance on the MCAS, and defined classification as the total number of students missclassified in any of the five groups divided by the number of students in all five groups. We found that the five variables selected by the MAD-based procedures described in this paper, with one addition, seconds on correct scaffolds, were also optimal when using achievement level classification error as the stage 2 criterion.

In future work, we plan to compare our algorithm to other online learning algorithms that can model data changes over time. It would be interesting to see if the learning curves would behave simlarly with these algorithms and how well they could do at predicting end-of-year assessment scores from monthly tutoring records. In particular we will consider comparing the performance of this method to the two models based on item-response theory of (Ayers and Junker 2006).

Acknowledgements

This research was made possible by the US Dept of Education, Institute of Education Science, "Effective Mathematics Education Research" program grant #R305K03140, the Office of Naval Research grant #N00014-03-1-0221, NSF CAREER award to Neil Heffernan, and the Spencer Foundation. All the opinions in this article are those of the authors, and not those of any of the funders.

This work would not have been possible without the assistance of the 2004-2005 WPI/CMU ASSISTment Team including Kenneth Koedinger, Elizabeth Ayers, Andrea Knight, Meghan Myers, Carolyn Rose all at Carnegie Mellon, Steven Ritter at Carnegie Learning, Neil Heffernan, Mingyu Feng, Tom Livak, Abraao Lourenco, Michael Macasek, Goss Nuzzo-Jones, Kai Rasmussen, Leena Razzaq, Terrence Turner, Ruta Upalekar, and Jason Walonoski all at WPI.

References

Ayers, E.; and Junker, B. W. 2006. Do skills combine additively to predict task difficulty in eigth-grade mathematics? Accepted, American Association for Artificial Intelligence Workshop on Educational Data Mining (AAAI-06), July 17, 2006, Boston, MA.

Beck, J. E.; Jia, P.; and Mostow, J. 2004. Automatically assessing oral reading fluency in a computer tutor that listens. *Technology, Instruction, Cognition and Learning* 2: 61-81.

Campione, J. C.; Brown, A. L.; and Bryant, N. R. 1985. Human abilities: An information-processing approach. *Individual differences in learning and memory*:103-126.

Feng, M.; Heffernan, N. T.; and Koedinger, K. R. 2006. Predicting State Test Scores Better with Intelligent Tutoring Systems: Developing Metrics to Measure Assistance Required. Accepted, 2006 Intelligent Tutoring Systems Conference, Taipei, Taiwan.

Feng, M.; Heffernan, N. T.; and Koedinger, K. R. in press. Addressing the Testing Challenge with a Web-Based E-Assessment System that Tutors as it Assesses. The 15th International World Wide Web Conference, Edinburgh, Scotland.

Feng, M.; Heffernan, N. T.; Mani, M.; and Heffernan, C. L. 2006. Does using finer-grained skill models lead to better predictions of state test scores?. Submitted to the International Conference on Intelligent Tutoring Systems (ITS 2006): Education Data Mining Workshop, Jhongli, Taiwan.

Hamilton, L. 1992. *Regression With Graphics: A Second Course in Applied Statistics.* Belmont, California.: Duxbury Press,

Heffernan, N.T., Koedinger, K.R., Junker, B.W. (2001). Using Web-Based Cognitive Assessment Systems for Predicting Student Performance on State Exams. Technical Report, Institute of Educational Statistics: US Dept. of Edcuation. Dept. of Computer Science, Worcester Polytechnic Institute Univ.

Koedinger, K. R.; Anderson, J. R.; Hadley, W. H.; and Mark, M. A. 1997. Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education* 8:30-43.

Koedinger, K. R.; Corbett, A. T.; Ritter, S.; and Shapiro, L. J. 2000. Carnegie Learning's Cognitive Tutor: Summary research results. White Paper. Technical Report, Carnegie Learning, Pittsburgh, PA.

Lee, C.; Ting, D. 2005. Predicting MCAS Scores from the ASSISTment System. Technical Report, Dept. of Statistics, Carnegie Mellon Univ.

Olson, L. 2005. State Test Programs Mushroom as NCLB Mandate Kicks In. *Education Week,* Nov. 30: 10-14.

Pardos, Z. A.; Heffernan, N. T.; Anderson, B.; and Heffernan, C. L. 2006. Using fine-grained skill models to fit student performance with Bayesian networks. Submitted to the International Conference on Intelligent Tutoring Systems (ITS 2006): Education Data Mining Workshop, Jhongli, Taiwan.

Snee, R. D. 1977. Validation of regression models: methods and examples. *Technometrics* 19: 415-428.

Shao, J. 1993. Linear model selection by cross-validation. *Journal of the American Statistical Association* 88: 486-494.

Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference.* New York, NY.: Springer-Verlag.

[2] http://www.doe.mass.edu/sda/ayp/about.html?section=3