Paper #2
Abstract Title Page

Title: Using Machine-Learned Detectors to Assess and Predict Students' Inquiry Performance

Author(s): Janice D. Gobert, Ryan Baker, & Michael Sao Pedro
Affiliation of all authors: Department of Social Sciences and Policy Studies, Worcester Polytechnic Institute, 100 Institute Rd., Worcester, MA 01609
Email addresses for all authors: {jgobert, rsbaker, mikesp}@wpi.edu
Contact email for paper: jgobert@wpi.edu

Background / Context:
In accordance with the National frameworks for inquiry (NRC, 1996), The Science Assistments project (www.scienceassistments.org; NSF-DRL# 0733286; NSF-DGE# 0742503, NSF-DRL# 1008649; U.S. Dept of Ed.# R305A090170Gobert et al, 2001, 2009) has developed a rigorous, technology-based learning environment that assists and assesses (hence, "assistments") middle school students in Earth, Life, and Physical Science so that teachers can assess their students' skills rigorously, frequently, and in the context in which they are developing, namely, during instruction (Mislevy et al, 2002). Our program of work represents a significant advance over other programs that utilize pencil and paper assessments because ours makes use of a state-of-the art logging infrastructure to do web-based tutoring and assessment (Razzaq et al, 2005). Our learning environment Science Assistments (www.scienceassistments.org;) scaffolds middle school students' scientific inquiry skills, namely, hypothesizing, designing and conducting experiments, interpreting data, warranting claims with evidence, and communicating findings.

Purpose / Objective / Research Question / Focus of Study:
We present work towards automatically assessing data collection behaviors as middle school students engage in inquiry within a physics microworld. In this study, we used machine learned models that can detect when students test their articulated hypotheses, design controlled experiments, and engage in planning behaviors using our inquiry support tools. We compared two approaches, an averaging-based method that assumes static skill level and Bayesian Knowledge Tracing, on their efficacy at predicting skill before a student engages in an inquiry activity and on predicting performance on a paper-style multiple choice test of inquiry and a transfer task requiring data collection skills.

Setting:
Our data were collected in a rural town in Central Massachusetts.

Population / Participants / Subjects: Participants were 134 eighth grade students, ranging in age from 12-14 years, from a public middle school in Central Massachusetts. Students belonged to one of six class sections and had one of two science teachers. Approximately 25% of the students are on free- or assisted-lunch and approximately 51% are "Below proficient" on the MCAS science test.

Intervention / Program / Practice:
Our learning environment Science Assistments scaffolds middle school students' scientific inquiry skills, namely, hypothesizing, designing and conducting experiments, interpreting data, warranting claims with evidence, and communicating findings. In this study, we present data from our state change microworld in which students engaged in inquiry to address the effects of various independent variables on a set of dependent variables. The variables in the microworld design were chosen to align with the Massachusetts curricular frameworks for middle school science. The specific tasks of interest in this study were testing their articulated hypotheses, designing and conducting experiments (using cvs), and planning experiments using the table tool.

Research Design:
Our machine learned models were trained using labels generated through a new method of manually hand-coding log files, called "text replay tagging" (Baker, Corbett & Wagner, 2006). We integrated these models into two approaches that assess latent student skill at runtime, Bayesian Knowledge-Tracing and an averaging approach that assumes static skill. Our approach resulted in behavior detectors that can accurately identify our inquiry behaviors of interest under student-level cross-validation. In addition, these models can be applied at run-time to drive intervention or assessment by the learning environment. We integrated these models into two approaches that assess latent student skill at runtime, Bayesian Knowledge-Tracing (BKT) and an averaging approach that assumes static skill.

Data Collection and Analysis:
The state change task (one of a set of domains in which we have used machine learning for assessment), which was used as an assessment of students' inquiry skills and content knowledge of the domain, engaged students in a series of inquiry tasks; the following were used in our analyses: testing articulated hypotheses, designing and conducting experiments (cvs), planning using the table tool.

Findings / Results:
We compared how well the averaging-based approach and the BKT model each fit actual student performance by predicting student inquiry skill using the probability that the student knew each skill of interest before each practice attempt. For each inquiry skill, the two models, the average-based model and BKT model, fit comparably well to student performance. There was at most a 2% difference between the two models for each behavior, testing hypotheses (BKT model A'=.79 vs. Avg model A'=.77), designing controlled experiments (A'=.74 for both), and planning using the table tool (BKT model A'=.70 vs. Avg model A'=.71). See table 1 for means and standard deviations for each approach.

The models for estimating proficiency also enabled us to determine if performance on authentic inquiry skills in the learning environment predicted performance on the transfer tests. Overall, each model for an authentic skill was correlated to its corresponding transfer test (when such a test was available), though somewhat unexpectedly; the average-based model predicted as well or better than the BKT model. Although the correlations were relatively weak, our findings provide some evidence that the skills for successfully engaging in authentic inquiry and answering equivalent paper test-style multiple-choice questions are related. See table 2 for correlations.

Conclusions:
Our findings provide some evidence that the skills for successfully engaging in authentic inquiry and answering equivalent paper test-style questions are related (Black, 1999; Pellegrino, 2001). Furthermore, our findings support the notion that authentic skill learned in one context can be applied to other domains, as shown by the significant correlation between performance in designing controlled experiments in two domains. As such, these models have considerable potential to enable future "discovery with models" analyses (cf. Baker, 2010) that can shed light on the relationship between a student's mastery of systematic experimentation strategies and their domain learning. Additional research will be needed to determine if these findings are

robust over different student populations and if the feature set and associated detectors are general enough (cf. Ghazarian & Noorhosseini, 2010) to be applicable to microworlds in other scientific domains. It will also be important to determine if these relationships will hold after incorporating scaffolding, thus giving students a better opportunity to both perform well despite incomplete knowledge, and to acquire these skills while using the microworld.

In summary, our work represents an advancement in automatically assessing and estimating procedural skills in an ill-defined domain, particularly the procedures that students use as they engage in scientific inquiry within a science microworld. Furthermore, this work represents the development of the first verifiable machine-learned models of inquiry behavior derived from student actions and human classifications. We also showed how these behavior detectors could be aggregated to form estimates of skill proficiency. These estimates enabled us to study the relationship between authentic inquiry skill and performance on multiple-choice test-based measures of inquiry, with implications for the future design of standardized tests in this domain. In addition, this paper presents iterative improvement to the development of machine-learned models of skill in ill-defined domains. Our text replay-tagging approach could easily be applied to generate validated detectors of procedural skill in other domains. Moreover, such detectors can be used as a basis for formulating estimates of skill to support scaffolding and prediction of external measures of learning.

Appendices A

References

Baker, R., Corbett, A., & Wagner, A. (2006). Human Classification of Low-Fidelity Replays of Student Actions. *Proceedings of the Educational Data Mining Workshop at the 8th International Conference on Intelligent Tutoring Systems*, (pp. 29-36).

Baker, R.S.J.d. (2010) Data Mining for Education. In McGaw, B., Peterson, P., Baker, E. (Eds.) *International Encyclopedia of Education (3rd edition)*, vol. 7, pp. 112-118. Oxford, UK: Elsevier.

Black, P. (1999). *Testing: Friend or Foe? Theory and Practice of Assessment and Testing*. New York, NY: Falmer Press.

Ferguson, G. (1976). *Statistical Analyis in Psychology and Education, 4th Edition.* McGraw-Hill Inc.

Ghazarian, A., & Noorhosseini, S. M. (2010). Automatic Detection of Users' Skill Levels Using High-Frequency User Interface Events. *The Journal of User Modeling and User-Adapted Interaction , 20* (2), 109-146.

Gobert, J.; Heffernan, N.; Koedinger, K.; Beck, J. (2009). *ASSISTments Meets Science Learning*. (AMSL). Proposal (R305A090170) funded by the U.S. Dept. of Education.

Gobert, J.; Heffernan, N.; Ruiz, C.; Kim, R. (2007). *AMI: ASSISTments Meets Inquiry*. Proposal NSF-DRL# 0733286 funded by the National Science Foundation.

Mislevy, R. J., Chudowsky, N., Draney, K., Fried, R., Gaffney, T., and Haertel, G. (2002). *Design patterns for assessing science inquiry*. Unpublished manuscript, Washington, D.C.

NSES. (1996). *National Committee on Science Education Standards and Assessment. (1996).* National Science Education Standards, Washington, D.C., National Academy Press.

Pellegrino, J. (2001). *Rethinking and redesigning educational assessment: Preschool through postsecondary.* Denver, CO: Education Commission of the States.

Razzaq, L., Feng, M., Nuzzo-Jones, G., Heffernan, N.T., Koedinger, K. R., Junker, B., Ritter, S., Knight, A., Aniszczyk, C., Choksey, S., Livak, T., Mercado, E., Turner, T.E., Upalekar. R, Walonoski, J.A., Macasek. M.A., and Rasmussen, K.P. (2005). The Assistment Project: Blending Assessment and Assisting. In C.K. Looi, G. McCalla, B. Bredeweg, and J. Breuker (Eds.) *Proceedings of the 12th Artificial Intelligence In Education*, Amsterdam: ISO Press. Pp. 555-562.

Appendices B

Table 1. Means and Standard Deviations for estimates of authentic skill and posttest measures, *N=134*.

	MAX	***M***	**SD**
AVERAGE-BASED MODELS			
Controlled experiments average		0.25	0.26
Testing hypotheses average		0.30	0.29
Planning using table tool average		0.05	0.11
BAYESIAN KNOWLEDGE TRACING MODELS			
Controlled experiments estimate		0.32	0.38
Testing hypotheses estimate		0.43	0.41
Planning using table tool estimate		0.01	0.08
DEPENDENT MEASURES			
Inquiry post-test: testing hypotheses	6	2.06	1.53
Inquiry post-test: cvs	4	2.09	1.18
Ramp transfer: cvs	4	1.59	1.70

Table 2. Correlations between post-test measures and each model's estimate of authentic skill, *N* = 134.

	Avg-based Model *r*	**BKT model *r***	***t* difference**
DEPENDENT MEASURES			
Inquiry post-test: testing hypotheses	.41 ***	.31 ***	2.50 *
Inquiry post-test: cvs	.26**	.26**	0.03
Ramp transfer: cvs	.38***	.37***	0.21

Note: The t difference between model correlations was computed using a significance of the difference between two correlation coefficients for correlated samples (Ferguson, 1976, pp.171-172).

$*p < .05$; $**p < .01$; $***\ p < .001$