Abstract Title Page

Not included in page count.

Title: Using model-tracing to conduct performance assessment of students' inquiry skills within a microworld.

Author(s): Janice D. Gobert & Kenneth R. Koedinger

Affiliation of all authors: Gobert: Worcester Polytechnic Institute; Koedinger, Carnegie Mellon University

Email addresses for all authors: jgobert@wpi.edu; koedinger@cs.cmu.edu

Contact email for paper: Janice Gobert, **jgobert@wpi.edu**

Background / Context:
Description of prior research and its intellectual context.

The National frameworks for science emphasize inquiry skills (NRC, 1996), however, in typical classroom practice, science learning often focuses on rote learning in part because science process skills are difficult to assess (Fadel, Honey, & Pasnick, 2007) and rote knowledge is prioritized on high-stakes tests. Short answer assessments of inquiry have been used (cf., Alonzo & Aschbacher, 2004; Songer, 2006), however, these tend to not align well to current national frameworks (Quellmalz, Kreikemeier, DeBarger, & Haertel, 2007) and it is unclear whether they properly identify inquiry skills (Black, 1999; Pellegrino, 2001). Hands-on performance assessments are more authentic (Baxter and Shavelson 1994; Ruiz-Primo & Shavelson, 1996), however, these are seldom used in schools because of difficulty with reliable administration and the resulting high cost.

The Science Assistments project (www.scienceassistments.org) 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 during instruction--in the context in which they are developing (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 (Razzaq et al, 2005).

Purpose / Objective / Research Question / Focus of Study: As a proof of concept for automated assessment of scientific inquiry skills, we used model-tracing (Corbett & Anderson, 1995; Koedinger & Corbett, 2006) to develop a cognitive model of science inquiry skills, particularly, the control for variables strategy (Chen & Klahr, 1999) and warranting claims with data. This model provides a rich qualitative, process-oriented scoring of students' inquiry "moves" within a guided scientific inquiry simulation for the domain of state change. We address the validity of this automated approach to performance assessment both quantitatively, in terms of reliability and predictive validity, and qualitatively, in terms of providing rich traces of student inquiry steps and "mis-steps" or haphazard inquiry (Buckley, Gobert et al, 2010).

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

Population / Participants / Subjects:
Participants. Participants were 78 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 (www.scienceassistments.org; NSF-DRL# 0733286; NSF-DRL# 1008649; U.S. Dept of Ed.# R305A090170) scaffolds middle school students' scientific inquiry skills, namely, hypothesizing, designing and conducting experiments, interpreting data, warranting claims with evidence, and communicating findings. The state change task, which was used as an assessment of students' inquiry skills and content knowledge of the domain, first allowed for student exploration in order to orient the learner to the interface and microworld. The students engaged in the next series of tasks; our data for the present study is drawn from these tasks.

1) “Try to find out how the size of the container (trial 1), amount of substance (trial 2), level of heat (trial 3), and cover status (trial 4) affects each of the dependent variables: the melting point of the ice, the time it takes the ice to completely melt, the boiling point of the water, and, the time it takes for the water to completely boil.
2) After each trial, students were asked to derive a conclusion from their data and select trials that supported their claim.
3) The *communicate your findings task*: “Pretend you are explaining your conclusions about the effects of cover status on each of the dependent variables to a friend who did not do the experiments. Discuss how you conducted the experiments and how you came to your conclusions. Be as specific as possible.”

Building upon past work (Koedinger, Suthers, & Forbus, 1997), we created a computational model of scientific inquiry with production rules that could trace the students’ moves in the simulation relative to an ideal model of scientific inquiry. In particular, the model tracked whether students’ initial hypotheses were scientifically accurate, whether the experimental trials they ran were relevant to their hypotheses, whether their trials used the control for variables strategy, whether their final analysis entered was supported or unsupported by their data, and whether they had collected appropriate experimental evidence that supported their final conclusion (relevant controlled trials).

Research Design:
We describe a proof-of-concept for performance assessment of students’ inquiry skills within a science microworld and present Cronbach’s alphas as reliability measures for each of our variables of interest. We will (in the full paper) also describe how our model-tracing method can be used to detect cases of confirmation bias and to detect cases of genuine discovery in students who are conducting scientific inquiry.

Materials. Pre- and post-tests for inquiry skills. A short battery of multiple-choice items (n=12) was used to get a baseline measure of their inquiry skills including hypothesizing, independent and dependent variables, the control of variables strategy, and data interpretation.
Domain Pre and Post Tests: A short battery of multiple-choice items for content knowledge (n=7) was used to get a baseline measure content knowledge of this domain.
Phase Change Microworld Activity. This microworld (described above) was developed to address the “phase change” related strands of the Massachusetts Frameworks for Physical Science at the middle school level.

Procedure. Pre- and post-tests for inquiry skills were administered before and after the students’ use of the phase change inquiry microworld.

The domain pre and post-tests were administered before and after the students’ use of the phase change inquiry microworld.

Data Collection and Analysis:
Within the Science Assistments system, all students’ inquiry actions are logged, thus students’ experimental trials for collecting data are automatically collected when the student hits the “run” button within the state change microworld. By applying model-tracing, as described above, to students’ log data we coded for the following variables: 1) CVS-relevant for each of the four

trials (whether a set of trials is using CVS and whether they are relevant to the student's articulated hypothesis), 2) tested-and-true for each of the four trials (whether a claim is supported based on data as collected by the student), and 3) lastly, an average of these scores, referred to as %cvs+true-tested for each of the four trials.

Findings / Results:

First, using data from our model-tracer, we calculated Cronbach's alpha for our variables of interest in order to ascertain the reliability across the 4 trials on each of the measures. The cronbach's alpha for the 4 CVS-relevant scores was 0.682, indicating an acceptable degree of internal consistency amongst the 4 measures for CVS-relevant; this suggests that we are getting consistency on our performance assessment for CVS-relevant across the four trials in terms of capturing when the student is conducting relevant hypotheses using the control of variables strategy (CVS). Secondly, the Cronbach's alpha for the 4 true-tested scores was 0.762, indicating a fairly high degree of internal consistency amongst the 4 measures for "tested-and-true" hypotheses; this suggests that we are getting consistency on our performance assessment in terms of capturing when the student has tested a hypothesis that is scientifically accurate. Lastly, the Cronbach's alpha for the aggregate of the two inquiry scores across the 4 trials, %CVS+true-tested, was 0.784, indicating a high degree of internal consistency amongst the measures. This suggests that we are getting a high degree of consistency on our performance assessment for the aggregate measure of CVS-relevant and tested-and-true hypotheses.

Secondly, we calculated correlations between our auto-scored performance measures of inquiry for CVS-relevant hypotheses with specific post-test inquiry items that should be, in theory, related. We obtained moderate correlations between our performance measures of inquiry and our post-test items for identifying an independent variable, identifying a dependent variable, and demonstrating the control of variables strategy (CVS). See Table 1 below.

Lastly, we used our model tracer to identify two interesting patterns of scientific inquiry: 1) when students engage in confirmation bias in their inquiry, even in the face of opposing evidence, and 2) when students make a discovery, using a controlled experiment to change an original false belief. This analysis focused on one of the four trials. Many students (45%) did not engage in repeated experiments. Of the remaining 79, the model identified 10 students who engaged in confirmation bias, and 8 students who made a genuine discovery.

Conclusions:

In this paper we have shown that we can use model-tracing as a method of performance assessment for science inquiry skills, an ill-defined domain. This builds upon the extensive work that has been done to date for well-defined domains such as math (Corbett & Anderson, 1995; Koedinger & Corbett, 2006). Additionally: 1) the reliability of our machine-scored measures of inquiry are highly consistent across the 4 Assistment activities or "trials", suggesting that we can reliably capture students' inquiry performance on these rich inquiry tasks, and 2) our measures are moderately correlated with post-test measures of inquiry performance for analogous concepts. Lastly, our data show that model-tracing can detect interesting patterns of student inquiry such as confirmation bias and overcoming confirmation basis. These are important data with respect to demonstrating auto-scoring of rich inquiry behaviors, but are also important, particularly the former, in terms of its implications for adaptive scaffolding of student inquiry, such as that being done by the Science Assistments group (www.scienceassistments.org; Gobert et al, 2007, 2009).

This work makes contribution to theoretical understanding of scientific inquiry, to its assessment, and to technical methods to auto-score inquiry. This represents an advance in this area since to date there has been difficulty in separating inquiry from the domain-specific context in which it was learned (Mislevy et al., 2002; Gobert, Pallant, & Daniels, 2010), and difficulty measuring inquiry skills due to their complexity and the amount of data required for reliable measurement (Shavelson et al, 1999).

References

Alonzo, A. and Aschbacher, P.R. (2004). Value Added? *Long assessment of students' scientific inquiry skills*. Presented at the Annual Meeting of the American Educational Research Association. San Diego, CA, April.

Baxter, G. P., and Shavelson, R. J. (1994). Science performance assessments: Benchmarks and surrogates. *International Journal of Educational Research, 21*, 279-298.

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

Buckley, B., Gobert, J., Horwitz, P., & O'Dwyer, L. (2010). Looking Inside the Black Box: Assessments and Decision-making in BioLogica. *International Journal of Learning Technology, 5* (2), 166-190.

Chen, Z., & Klahr, D. (1999). All Other Things Being Equal: Acquisition and Transfer of the Control of Variables Strategy. *Child Development, 70* (5), 1098-1120.

Corbett, A., & Anderson, J. (1995). Knowledge-Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction, 4*, 253-278.

Fadel, C., Honey, M., and Pasnick, S. (2007). Assessment in the Age of Innovation, *Education Week, Volume 26 (38),* 34-40.

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.

Gobert, J.D, Pallant, A.R., & Daniels, J.T.M. (2010). Unpacking inquiry skills from content knowledge in Geoscience: A research perspective with implications for assessment design. *International Journal of Learning Technologies, 5(3),* 310-334.

Koedinger, K., & Corbett, A. (2006). Cognitive Tutors: Technology Bringing Learning Sciences to the Classroom. In R. Sawyer, *The Cambridge Handbook of the Learning Sciences* (pp. 61-77). New York, NY: Cambridge University Press.

Koedinger, K. R., Suthers, D. D., & Forbus, K. D. (1999). Component-based construction of a science learning space. *International Journal of Artificial Intelligence in Education, 10, NB need page numbers*

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.

Quellmalz, E., Kreikemeier, P., DeBarger, A. H., and Haertel, G. (2007). *A study of the alignment of the NAEP, TIMSS, and New Standards Science Assessments with the inquiry abilities in the National Science Education Standards*. Presented at the Annual Meeting of the American Educational Research Association, April 9-13, Chicago, IL.

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.

Ruiz-Primo, M. A., and Shavelson, R. J. (1996). Rhetoric and reality in science performance assessments: An update. *Journal of Research in Science Teaching, 33*(10), 1045-1063.

Appendix B. Tables and Figures

Table 1. Correlations between performance measures of inquiry and multiple-choice post-test measures of inquiry.

	CVS-relevant	test-and-true	%CVS+true-tested
Inquiry Posttest: Testing Hypotheses	0.418	0.455	0.485
Inquiry Posttest: Controlled experiments	0.372	0.304	0.376
Ramp Transfer: Controlled experiments	0.407	0.347	0.420