Murray, T., Rath, K., Woolf, B.P., Marshall, D., Bruno, M., Dragon, T., Kohler, K., Evaluating Inquiry Learning Through Recognition-Based Tasks. Proceedings of the 12th International Conference on AI and Education, July 2005, Amsterdam.

Evaluating Inquiry Learning Through Recognition-Based Tasks

Tom Murray, Kenneth Rath[†] , Beverly Woolf, David Marshall,
Merle Bruno[††], Toby Dragon, Kevin Kohler, Matthew Mattingly

University of Massachusetts, Amherst, MA
[†] Peterfreund Associates, Amherst, MA
[††] Hampshire College, Amherst, MA
contact: tmurray@cs.umass.edu

Abstract. The Rashi inquiry learning environment for human biology was evaluated using a new instrument for assessing gains in scientific inquiry skills. The instrument was designed to be sensitive to the small pre-post skill gains that are hypothesized for short learning interventions. It is also designed to be scored with less effort than the verbal protocol analysis methods most often used to asses higher order skills. To achieve these ends the instrument is "item-based", "recognition-based" and "difference-based." We describe our assessment design method and results of its first use.

1. Introduction

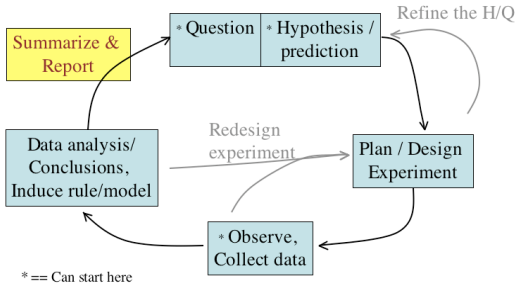
Rashi is a domain independent architecture for inquiry learning environments. It contains tools that allow learners to gather data, pose multiple hypotheses, and create arguments that support hypotheses by linking to supporting or refuting data. We are using Rashi to build inquiry learning environments in human biology, geology, and forest ecology, all for undergraduate level science. Though inquiry skills, like all higher order thinking skills, are difficult to assess [1], it is important that we develop methods for assessing these skills because they are essential in many types of work and problem solving, and they are given high priority in many educational standards and frameworks.

A common problem in research into advanced learning environments is that the software is not able to be tested in authentic contexts over extended periods of use. Such systems usually have significant pedagogical "depth" but little content scope, and when they are employed in classrooms their content applies to a very small portion of the curriculum. Also, it may be difficult to find instructors willing to "give up" significant course time to an alternative approach. The fact that our interventions may be limited to weeks or even hours is at odds with the slow rate of improvement expected for higher order cognitive skills. In order to evaluate these interventions instruments need to be sensitive to small learning gains.

In this paper we describe our first attempts with a new methodology for developing assessments for inquiry learning environments. Our goals are to design inquiry assessment instruments that are: 1) sensitive to small changes in skill level, and b) less labor intensive than most currently used methods. The method uses recognition-based (as opposed to recall), item-based (as opposed to free-form), and difference-based tasks (as described later). We describe our first use of this method, its results, and planned improvements on the method. Data analysis of the results revealed no statistically significant conclusions, and this we attribute to a non-optimal subject context (there was insignificant motivation for the volunteers to take the task seriously) which will be avoided in future trials. Thus the contribution of this paper is more in the description and discussion of the methodology than about evaluation results.

2. An Inquiry Learning Model

Our model of inquiry learning is based in part on knowledge from the experts we work with, and in part on the inquiry-based learning literature ([2],[3],[4], among others). Figure 1 shows our model of the "scientific inquiry process," which combines elements from other models.



The following inquiry skills have been identified as most important by our subject-matter experts:

Table 1

1.	**Understand** the task and what constitutes completion of the task
2.	**Differentiate** observation (and data) from inferences
3.	**Justify** hypotheses with arguments
4.	**Explain** inferences and hypotheses
5.	**Explore** observation, measurement, and information resources
6.	**Cite** source documents
7.	**Systematically** gather, interpret, and organize information.
8.	**Communicate** a clear summary of your findings in written form.

We have used this skill list to inform the design of the Rashi tools, and to inform the design of our evaluations.

3. Assessment design issues

We will describe three methodological decisions which resulted in our assessment task being "item-based," "recognition-based," and "difference-based."

Item-based tasks. The most common methods used in researching inquiry-based, discovery-based, or open-ended learning environments are qualitative and ethnographic methods. Such methods include analysis of verbal data from peer work or structured interviews and analysis of written work from assignment portfolios or journals. They are appropriate for interpretive research aiming for a "thick" characterization of the student/user experience for a small number of cases, and are very labor-intensive. The literature includes many examples of inquiry-based educational technology projects that have used such evaluation methods ([2],[5],[6],[7],[8],[9]). In exploratory research where the key questions and constructs are still being worked out, ethnographic methods are used because they allow the nature of the data analysis to evolve during the analysis. But when a theoretical framework already exists, more specific types of tasks can be designed. These "closed" tasks, such as sorting, ranking, and comparison tasks (see [10],[11]), can be more reliable and generalizable, but tend to have less authenticity, which may effect the ecological validity of results.

As a compromise between closed tasks and more open-ended tasks, case data (verbal or written) should have clear segmental boundaries such as answers to questions or problem solution steps, allowing many data points per subject (see "intra-sample statistical analysis" in [12]). We will call this approach item-based to refer to segmenting the task into discrete task items. Scoring rubrics can then address more constrained tasks ([13],[14]).

Recognition-based tasks. As mentioned, significant gains in higher-order thinking skills usually require significant learning time. Yet for these trials we were limited to 2 or 3 sessions of 2 to 3 hours each. Given these constraints, we hypothesized that a recognition task would be more likely to show skill improvement than a recall task. Recognition learning usually occurs more easily than recall learning. For example, imagine that someone reads a list to you and then reads from another list and asks whether or not each of the items was on the original list. This task is easier than trying to recall all of the items from the original list without being given any cues. Our recognition task for inquiry learning skills involves rating the quality of a hypothetical problem solution, rather than generating a problem solution from scratch.

Roth & Roychoudhury [3] discuss the importance of a socio-cultural perspective on teaching scientific inquiry, noting that "new and more powerful skills and concepts can be observed in social interaction long before they are exhibited by individuals" pg (133). The collaborative context can be particularly useful to exposing learning gains. We propose that this will still be the case (thought to a lesser degree) in "mock" sociological contexts, as in when a student is asked to critique or rate the work of a hypothetical peer.

Difference-based tasks. The third methodological decision was that the post-test task involved asking subjects to *improve* upon their pre-test answers as opposed to solving an entirely new problem. We believe that this "difference-based task" will further sensitize the instrument to small gains in inquiry skills. It also removes some of the variability introduced when using different pre-and post test tasks that have not undergone rigorous psychometric verification of equivalence.

Before further describing the instrument we will briefly describe the software evaluated.

4. Description of Biology Domain and Rashi Tools

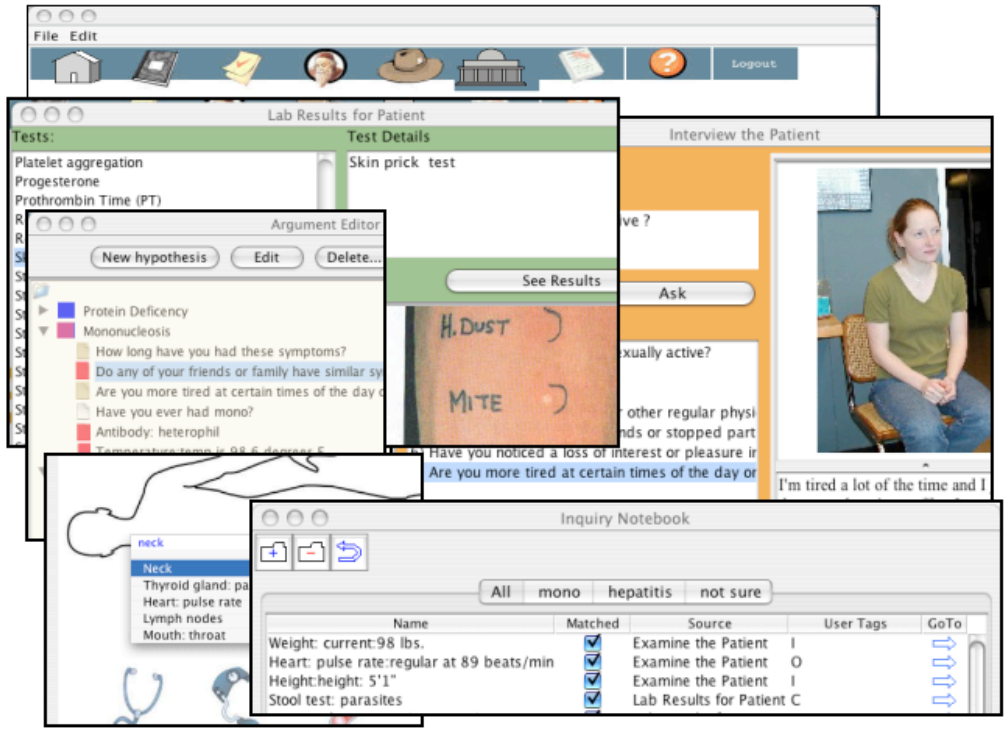
Rashi domains and inquiry tasks. In the Rashi Human Biology (HB) Tutor, learners are presented with medical cases and attempt to diagnose the patient's condition. They can interview the patient, perform various physical exams, and order lab tests for blood, urine, etc. Rashi includes an authoring tool [15] that allows authors to create new cases. The Rashi HB tutor is based on a case-based classroom teaching method used by one of the co-authors at Hampshire College [16]. Eight cases have been authored from Rashi HB, based on medical conditions including mold allergies, hyperthyroidism, and lactose intolerance.

Next we give a very brief overview of the tools available to learners in the Rashi system (and see [17],[18],[19],). Rashi provides a set of tools that map onto the inquiry skills mentioned in Table 1 .

- **Case Orientation Screen**: Provides information about the case and general problem solving instructions. (Supports skill #1 in Table 1.)
- **Data gathering tools**: Each domain has its own set of data gathering tools. For the human biology domain they include a patient interview, physical exam, and lab tests. (Supports skill #5)
- **Inquiry Notebook**: Gathered data is saved to the inquiry notebook, which allows the setting of the data source, confidence level, and data type (hypothesis, measurement, observation, etc.) for each item. Data can be organized into folders (like having different pages in a research notebook), and keyword tags can be entered for sorting the items. (Supports skills #2, 6, 7)
- **Argument Editor**: Users create hypotheses and create arguments for and against them through links to notebook data items. Hypotheses are rated (e.g. top, possible, ruled out), and the argument relationship types are specified (e.g. supports, refutes, etc.).

Users can enter explanations for their hypotheses and for each argument link. (Supports skills #3, 4)

Rashi includes a Planning Scratch Pad (for skill #7), a Sources Editor (for skill #6), a Concept Library (for skill #5), and a Reporting Tool (for skill #8). The figure below shows some of the tools from the Biology domain (lab test results in upper left; patient interview in upper right, physical exam in lower left, argument editor in middle left, and notebook in lower right, with the main screen showing icons to access the tools shown at the very top). Rashi also has an intelligent coach, but this was turned off for these studies because the advice it gives was not yet robust enough. Also, we wanted this study to serve as a baseline for evaluating the system with the coaching turned on.



5. Methodology

Evaluation context and goals.

In the Fall of 2004, we evaluated inquiry skill gains resulting from Rashi HB use in two college classrooms. The first trial was in Bruno's small introductory Biology class, and served as a pilot test of our inquiry skills instrument. The second marked the first time Rashi had been used in the context of a large lecture class.

Having developed and tested Rashi in the context of a small-sized classroom with a teacher skilled in case-based inquiry pedagogy (which is not the context in which we expect it to show the largest benefit over the usual classroom experience), we wanted to test the system in the context of a larger classroom where the instructor did not have a high level of inquiry teaching skill.

Unfortunately, we were unable to find a large-sized college class in Fall of 2004 where the Rashi activities could be integrated, but we found an large introductory biology class for which the Rashi activities could be assigned as *extra credit*. The amount of extra credit time available was limited to 6-8 hours, including, instructions, and survey/test-taking.

Evaluation instrument

Developing evaluation tasks and instruments for inquiry learning environments is still very much a "black art," so below we describe in some detail how we developed ours. The Rashi tools are designed with a specific inquiry task model in mind, and we designed our evaluation task according to this model. As mentioned, the evaluation task involved presenting the subject with a hypothetical (and "imperfect") case solution created by an imaginary "student investigator", and asking the subject to evaluate its quality.

A. **Task design**. We wanted the evaluation task structure to parallel the task structure of using the Rashi tools to solve a case, so be broke up the Hypothetical Case Solution into three parts roughly corresponding to the main Rashi Tools. Solution Part A ("Beginning the Case") consisted of lists or initial hypotheses and what information is needed to confirm or reject them. Part B ("Data Collection") consisted of a list of data collected, with reasons. Part C ("Diagnosis Justification") consisted of a final set of accepted and rejected hypotheses, with justifications pointing to the data collected.

For all three parts of the pre-test, the instructions said: "List at least two strengths and two weaknesses of the investigator's notes." For the post-test, subjects were given exactly the same exercise and a copy of their previous answers. The only difference was the instruction to look at their pre-test answers and list at least one additional strength and weakness of the investigator's notes.

B. **Ideal Solution Characteristics**. We developed a model solution rubric describing the characteristics of a "correct" set of investigator notes for the task. We developed this list from the list of inquiry skills and through piloting the instrument and looking at the types of correct and incorrect statements that students made.

C. **Case Creation**. We created a case that focused on a different medical topic than that used in the Rashi software. The Case Description given to subjects included: "Jean Rockford, a 26-year-old woman, comes to see you with a 6-month history of increasing nervousness, irritability, and heat intolerance...."

D. **Ideal Solution Instance**. We constructed an ideal diagnosis solution, including approximately 15 items for each of the three parts, which included all of the characteristics of an ideal solution.

E. **Imperfect Solution**. We modified this ideal solution to create the final Hypothetical Case Solution with errors of omission and commission. This was a delicate "operation" because we felt the final investigator notes should have a range of easy-to-notice to difficult-to-notice errors geared to differentiate skill levels. In addition, the entire set of Investigator's notes had to look reasonable, being mostly correct but with a tractable number of identifiable problems.

F. **Scoring Rubric Development**. Finally we developed a scoring rubric geared for the specific case. The "imperfect solution" had a total of 16 faults, for a total of 16 possible points in the "list the weaknesses" questions (the "strengths" questions were not scored).

Experimental method

Experimental and control groups. In addition to the Rashi-using experimental group, we had three additional comparison groups, named according to the task given: Non-interactive case investigation, Inquiry article reading task, and Biology article reading task. These groups were created to allow credit assignment for any gains observed in the Rashi-using group (i.e. to attribute such gains to the interactive software, or the case-based instructional method, or to an exposure to inquiry concepts).

The **Rashi** Group used the Rashi system to investigate a medical case. The **Non-interactive** Group was given the same medical case to diagnose, but instead of using the Rashi system for their investigation they used a web site with static information about the case and were given worksheets with tables for keeping track of "things I need to know,"

"data gathered" and "diagnostic hypotheses." Both the Rashi group and the Non-interactive-inquiry group were asked to write up a 1-3 page summary report of their investigation and conclusions, and email this to us. The **Inquiry-reading** Group was given an article about using inquiry learning methods in science, and the **Biology-reading** Group was given a research article on diet's relationship to cardiac illness. Both reading groups were asked to write 1-3 page summaries of the articles and email them to us.

We hypothesized that inquiry learning improvements in the four groups would be ordered as: Rashi > Non-interactive task > Inquiry-reading > Biology-reading. Our reasons were as follows. The more realistic and interactive features of Rashi, plus the tools it gives students to organize and visualize information, should have helped students focus on their inquiry process and thus improve skills, as compared with the non-interactive task. Constructivist learning theory predicts that the two inquiry tasks would fare better than the two reading tasks. Also, we expected that reading an article about inquiry learning might have a slight effect on students, while reading an article on an unrelated topic should not.

Additional measures

Software use records. Our software currently stores all student work on a central server, but does not record each student action as they are using the Rashi tools. For this study we compiled a number of feature-use statistics based on the final state of the subject's work

Attitude Survey. The students in the Rashi Group filled out a survey appended to the on-line post-test. The survey included a 11x3 response matrix where the 11 rows listed activities or skills that the software supports (e.g. understanding the entire inquiry process, gathering data and information, citing the sources of information) and the columns asked: A. "How *successful* were you at the following activities"; B. "How *easy* was it for you to do these activities"; and C. "How *important was Rashi* in your ability to do these activities." For each of the 33 cells in the response matrix, students selected from three Likert-scale values. In addition, subjects were asked how much time they spent on the Rashi task.

Experimental Context

Volunteers from an undergraduate biology class of about 500 students were offered extra credit for participating in the study. Of the 140 students who signed up and began the processes, only 74 finished all required tasks. The number of students who completed all tasks, along with average self-reported time, is shown below.

Group	N	Time to Complete
Rashi	17	2.4 hours
Non-interactive-inquiry	18	2.5 hours
Inquiry-reading	17	2.4 hours
Biology-reading	22	2.0 hours

6. Results and Analysis

The following table gives the average test scores, their difference, and the standard deviation of that difference, t-test, and significance for each of the four groups.

Group	N	Average Pre	Average Post	Average Diff	SD Diff	t-test	p
Rashi	17	0.71	1.00	0.29	0.47	2.58	0.020
Non-interactive	18	1.00	1.61	0.61	0.78	3.34	0.004
Inquiry-reading	17	0.94	1.29	0.35	0.61	2.40	0.029
Biology-reading	22	0.36	0.59	0.23	0.43	2.49	0.021
Total	**74**	**0.73**	**1.09**	**0.36**	**0.59**	**5.35**	**0.000**

The results indicate an extreme floor effect (with average pre and post tests scoring about 1 out of a possible 16 points). There was no significant differences between groups on any measure. An ANOVA analysis of the results found that there were no statistically significant differences in the amount of improvement on inquiry skills across the four groups ($F(3, 70) = 1.58$, $p = 0.20$). The effort given by students in all four groups is similar (2 to 2.5 hours), though we expected students in the two inquiry tasks to spend significantly more time than they did on the task. (Note: Because the of the difference-based nature of the post-task, we would expect all post-tests to have higher scores then pre-tests, thus the low p values.) Combining the first two groups into an "inquiry-based" set and the last two into a "reading-based" set and comparing inquiry-based with reading-based also shows no significant differences .

Attitude survey results. As in past formative evaluations of Rashi, the survey did not indicate any significant problems with the software. We interpret these results as supporting the usability of the software and its perceived usefulness, especially given the short amount of time students were introduced to it and used it, and the fact that the study task did not relate to their current classroom activities.

Software use metrics. Since there were no significant differences between the pre- and post tests, we will call the subject's pre-test score their "inquiry skill level." There were significant correlations between inquiry skill level and some of the Rashi use metrics. In particular, there were significant positive correlations between inquiry skill level and the number of hypotheses posed, the number of arguments, the number of items in the notebook, the number of explanations entered by students, the use of notebook organizing tools, and the overall use of Rashi tools. As this is what one would expect, this adds some credence to the ecological validity of the pre-post instrument.

7. Discussion of Results

Floor-effect. As mentioned, our evaluation suffered from a significant floor effect, which makes it difficult to compare results of the four experimental groups. Some of this can be attributed to the design of the instrument, but we believe that mostly the floor effect is a result of characteristics of the subject population. We believe that the subjects were not motivated to take the study very seriously and put the necessary mental effort into the evaluation and intervention tasks. We believe that this was because: 1) the tasks were not integrated into the classroom experience and had nothing to do with content covered in the class; 2) volunteers signed up only to receive extra credit, and did not take the evaluation tasks very seriously because they were only required to complete the steps of the study to receive extra credit.

Improvements. We plan to carry out evaluations of Rashi in about 5 classrooms in 2005. Improvements based on lessons learned from the current study will include: 1) clearer pre-post test instructions to focus subjects on inquiry-specific skills; 2) rewording the "2 or more" strengths and weaknesses questions to encourage more answer items; 3) performing the study in classrooms that have the intervention activities more integrated into classroom activities.

8. Conclusions

This study did not yield very informative results due to floor effects, which in the future should be remedied by one or a combination of the improvements mentioned above. However, we believe that our suggestions for the development of assessment instruments are innovative in the context of assisting inquiry learning environments, and worth pursuing further.

To summarize, our goals were 1) to develop an instrument sensitive to changes in inquiry skills after relatively brief interventions, and 2) to develop an instrument that could

be scored with relatively little effort. We believe that we succeeded on the second point, since the scoring of all 74 pre and 74 post tests was done by one person within a single day.

Our methods for developing more sensitive instruments for inquiry skill included creating an assessment task that was "recognition-based," "item-based," and "difference-based," as described above. Due to the difficulties with the present study, we do not know yet whether these methods are in fact useful. Our further studies in 2005 will answer this question.

A further methodological innovation was that we used system tracking data along with skill assessment and survey data, which is rarely done in studies of inquiry learning systems. This allows us to construct more elaborate explanations for any significant differences we find within or between experimental groups. Our method of constructing a comparison task starting with ideal solution characteristics based on the inquiry model, then creating an ideal solution, and then perturbing the ideal solution to create the final imperfect Hypothetical Case Solution also seems unique to inquiry learning environment evaluations.

References

[1] Champagne, A.B., Kouba, V.L., & Hurley, M. (2000). Assessing Inquiry. In J. Minstrell & E. H. van Zee (Eds.) *Inquiry into Inquiry Learning and Teaching in Science*. American Association for the Advancement of Science, Washington, DC.

[2] White, B., Shimoda, T., Frederiksen, J. (1999). Enabling students to construct theories of collaborative inquiry and reflective learning: computer support for metacognitive development. *International J. of Artificial Intelligence in Education* Vol. 10, 151-1182.

[3] Roth, W. & Roychoudhury, A. (1993). The Development of Science Process Skills in Authentic Contexts. *J. of Research in Science Teaching*, Vol. 30, No 2. pp. 127-152.

[4] Edelson, D.C., D.N. Gordin, and P.D. Pea (1999). "Addressing the Challenges of Inquiry Based Learning Through Technology and Curriculum Design." The *Journal of Learning Sciences*. 8(3&4): 391-450. 1999..

[5] Azevedo, R., Verona, E., Cromley, J.G. (2001). Fostering learners collaborative problem solving with RiverWeb. J.D. Moore et. al. (Eds.) *Proceedings of Artificial Intelligence in Education,* pp. 166-172.

[6] Krajcik,J., Blumenfeld, P.C., Marx, R.W., Bass, K.M., Fredricks, J. (1998). "Inquiry in Project-Based Science Classrooms: Initial Attempts by Middle School Students" *J. of the Learning Sciences*, 7(3-4), pp 313-350, 1998.

[7] van Joolingen, W., & de Jong, T. (1996). Design and Implementation of Simulation Based Discovery Environments: The SMILSE Solution. *Jl. of Artificial Intelligence in Education* 7(3/4) p 253-276.

[8] Zachos, P., Hick, T., Doane, W., & Sargent, C. (2000). Setting Theoretical and Empirical Foundations for Assessing Scientific Inquiry and Discovery in Educational Programs. *J. of Research in Science Teaching* 37(9), 938-962.

[9] Murray, T., Winship, L. , Stillings, N. (2003B). Measuring Inquiry Cycles in Simulation-Based Leaning Environments. Proceedings of Cognitive Science, July, 2003, Boston, MA.

[10] Mestre J. P. (2000). Progress in Research: The interplay among theory, research questions, and measurement techniques. IN Lesh (ed)..

[11] Toth, E.E., Klahr, D, Chen, Z. (2000). Bridging research and practice: A cognitively based classroom intervention for teaching experimental skills to elementary school children. *Cognition and Instruction*, 18(4), 423-495.

[12] Shaffer, D.W. & Serlin R.C. (2005). What good are statistics that don't generalize? *Educational Researcher* 33(9), 14-25.

[13] Lunsford, E. & Melear, C. T. (2004). Using Scoring Rubrics to Evaluate Inquiry. J. of College Science Teaching 34(1), 34-38.

[14] Stillings, N. A., Ramirez, M. A., & Wenk, L. (1999). Assessing critical thinking in a student-active science curriculum. Paper presented at the meeting of the National Association of Research on Science Teaching, Boston, MA.

[15] Murray, T., Woolf, B. & Marshall, D. (2004). Lessons Learned from Authoring for Inquiry Learning: A tale of authoring tool evolution. J.C. Lester et al. (Eds.). ITS 2004 Proceedings, pp 197-206.

[16] Bruno, M.S. & Jarvis, C.D. (2001). It's Fun, But is it Science? Goals and Strategies in a Problem-Based Learning Course. J. of Mathematics and Science: Collaborative Explorations.

[17] Murray, T., Woolf, B., Marshall, D. (2003). Toward a Generic Architecture and Authoring Tools Supporting Inquiry Learning. Proceedings of AI-ED'2003, 11th World Conference on Artificial Intelligence in Education, 20-24 July, 2003. Sydney, pp. 488-490.

[18] Woolf, B.P., Marshall, D., Mattingly, M., Lewis, J. Wright, S. , Jellison. M., Murray, T. (2003). Tracking Student Propositions in an Inquiry System. Proceedings of AI-ED'2003, 11th World Conference on Artificial Intelligence in Education, 20-24 July, 2003. Sydney, pp. 21-28.

[19] Woolf, B.P, Murray, T., Marshall, D., Dragon, T., Kohler, K., Mattingly, M., Bruno, M., Murray, D. & Sammons, J. [in this volume]. Critical Thinking Environments for Science Education.