Abstract Title Page

Not included in page count.

Title: Rimac: A Natural-language Dialogue System that Engages Students in Deep Reasoning Dialogues about Physics

Author(s): Sandra Katz, Pamela Jordan, Diane Litman and the Rimac Project Team

Abstract Body

Limit 5 pages single spaced.

Background / Context:

Description of prior research and its intellectual context.

Recent studies have shown that U.S. students lag behind students in other developed countries in math and science (eg., Glod, 2007). Because one-on-one tutoring has proven to be a highly effective form of instruction (e.g., Bloom, 1984; Cohen, Kulik, & Kulik, 1982; Corbett, 2001), many educators and education policy makers have looked to intelligent tutoring systems (ITSs) as a means of providing cost-effective, individualized instruction to students that can improve their conceptual understanding of, and problem-solving skills in, math and science. However, even though many ITSs have been shown to be effective, they are still not as effective as human tutors (Corbett et al., 1999).

The goal of the Rimac Project is to take a step towards meeting President Obama's challenge to produce "learning software as effective as a personal tutor" (Obama, 2009, included in Hamblin, 2009). We are doing this by building an enhanced version of a natural-language dialogue system that engages students in deep-reasoning, reflective dialogues after they solve quantitative problems in *Andes*, an intelligent tutoring system for physics (VanLehn et al., 2005).[†] Enhancements to this system focus on addressing a key limitation of natural-language (NL) tutoring systems: although these systems are "interactive" in the sense that they try to elicit explanations from students instead of lecturing to them, automated tutors do not align their dialogue turns with those of the student to the same degree, and in the same ways, that human tutors do. In particular, automated tutors often fail to reuse parts of the student's dialogue turns in their own turns, to adjust the level of abstraction that the student is working from when the student is over-generalizing or missing important distinctions between concepts, and to abstract or further specify correct student input when doing so might enhance the student's understanding. The importance of "interactivity" or "verbal alignment" is supported by language processing theory (e.g., Pickering and Garrod's Interactive Alignment model; Pickering & Garrod, 2004), this research team's prior work demonstrating a positive relationship between abstraction/specialization relations during instructional dialogue and learning (e.g., Ward & Litman, 2008; Ward et al., 2009), and related work on classroom discourse (e.g., Michaels et al., 2008).

Purpose / Objective / Research Question / Focus of Study:

Description of the focus of the research.

The natural-language tutorial dialogue system that we are developing will allow us to focus on the nature of interactivity during tutoring as a malleable factor. Specifically, it will serve as a research platform for studies that manipulate the frequency and types of verbal alignment processes that take place during tutoring, such as abstraction and specification, and thereby enable us to test our prediction that the benefits of "interactivity" hinge on the nature, frequency,

[†] Rimac is the name of a river whose source is in the Andes. Its name is derived from the Quechua word, rimac, which means "talking;" hence the nickname for the Rimac: "talking river." We thus considered Rimac to be well-suited to a dialogue system embedded in the Andes tutoring system.

and appropriateness of its occurrence. The purpose of this poster is to introduce the Rimac project and system to the education research community. We have started the first year of a three-year IES development project.

Setting:
Description of the research location.

The University of Pittsburgh, a Pennsylvania-state related university, and several Pittsburgh urban and suburban high schools.

Population / Participants / Subjects:
Description of the participants in the study: who, how many, key features or characteristics.

The subject pool for these trials will be students enrolled in a first-year physics course at the University of Pittsburgh (Pitt), and high school students taking physics in several Pittsburgh urban and suburban schools. High school students are the target population for a pilot study in the final year of this project. Pitt students who achieved a B or better in a high school physics course will be recruited for field testing prior to that (i.e., during the system development phase).

Four physics teachers will be using the tutoring system during a pilot test of our dialogue-enhanced version of Andes that will take place during the third (final) year of the project. We are targeting the college-bound or "scholars" track of physics students in these schools—that is, the middle tier of students, not those in remedial or advanced placement math and science courses. We anticipate a sample size of 100 students, across the physics classes taught by participating high school teachers.

Intervention / Program / Practice:
Description of the intervention, program or practice, including details of administration and duration.

The intervention is the Andes physics tutoring system (e.g., VanLehn et al., 2005), enhanced to include the natural-language dialogue system that engages students in deep-reasoning conversations that we are developing, Rimac. A screen image of the Andes user interface is shown in Figure 1 (Appendix B). Andes was designed to be used as a "homework helper." It provides students with a graphical interface, immediate color-coded feedback on the student's vector diagrams and equations, and an online coach that can provide help on demand.

The reflective dialogues that we are developing are designed to address a limitation of the current (non-dialogue) version of Andes. Although students who used Andes as a "homework helper" in physics courses at the U.S. Naval Academy outperformed students in traditional (non-Andes) sections on tests of quantitative problem-solving skill (VanLehn et al., 2005), no such differences have been observed for conceptual knowledge and qualitative problem-solving ability. In addition, instructors who use Andes in their course frequently complain that the tutor needs to do a better job with this important, qualitative/conceptual aspect of physics instruction.

The particular intervention that we are developing and pilot testing is a set of natural-language, "reflective dialogues" that students will engage in after they solve problems in Andes. These dialogues are designed to present deep-reasoning questions about physics (Craig, Driscoll, & Gholson, 2006)—questions that invite students to think about the physics concepts and principles that are associated with just-solved problems. A sample reflective dialogue is shown in Figure 2, Appendix B. Our prior work with incorporating reflective dialogues in tutoring systems (in physics and other domains) yielded mixed results. For example, several studies demonstrated that college students who used an initial reflective dialogue-enhanced version of Andes significantly outperformed a no-dialogue (standard Andes) control group on an experimenter-designed test of conceptual knowledge. However, no significant differences between groups have been observed on a standard measure of conceptual knowledge and qualitative problem-solving ability in physics, similar to the Force Concept Inventory (Hestenes, Wells, & Swackhamer, 1992). Furthermore, comparisons of students who used our previous dialogue-enhanced version of Andes with students who used the standard version have failed to demonstrate transfer of conceptual knowledge to quantitative problem-solving ability (e.g., Connelly & Katz, 2009; Katz, Connelly, & Wilson, 2007). In short, we recognized the need for improvement in the reflective dialogue intervention that we added to Andes.

Our main approach to improving these reflective dialogues is to design them to align the conversational turns of the automated tutor with those of the student, in similar ways that we and other researchers have observed human tutors to do. For example, the automated tutor will further specify, or abstract over the student's dialogue turns—a form of interactivity that we have found to predict learning (e.g., Ward & Litman, 2008; Ward et al., 2009). One of our main challenges is to specify decision rules that can guide this automated abstraction/specialization process. These rules are being derived from analyses of live student-tutor reflective dialogues, and participating teachers' feedback on Rimac's performance, throughout several cycles of system development and field testing.

Although the standard, no-dialogue version of Andes has been used for several years in college-level physics courses, it has only recently been used in a few high school physics classes. Our project will test the feasibility of implementing Andes, enhanced with natural-language reflective dialogues, at the high school level, and pilot test the system to determine if students who use Andes+Rimac outperform students who use the standard version of Andes on conceptual and quantitative problem-solving measures. Pilot testing of the tutoring system will take place during 3-4 weeks of instruction on selected topics, in participating high school physics classes.

Andes is *freeware*, available online or as a standalone version that can be downloaded from http://www.andestutor.org/. Hence, if future efficacy trials of the enhanced, natural-language dialogue version of Andes that we are developing (Rimac) verify that this version of the system is more effective than the standard, no-dialogue version of Andes, the software would be readily available to high school teachers and students.

Research Design:

Description of research design (e.g., qualitative case study, quasi-experimental design, secondary analysis, analytic essay, randomized field trial).

Randomized field trials. We will have two conditions, intervention and control. Randomization will take place within each teacher's class.

Data Collection and Analysis:

Description of the methods for collecting and analyzing data.

The first two years of this project will be spent developing Rimac in close collaboration with high school physics teachers, in order to develop a robust set of reflective dialogues that are suitable for school students. In the final year of the project, we will pilot test Rimac in order to address the following questions: (1) do students who use the enhanced reflective dialogue version of Andes show greater gains from pretest to posttest than a control group of students who use the standard, no-dialogue version of this system, (2) for what types of knowledge do we find any observable differences in gain scores—that is, conceptual knowledge, and/or problem-solving ability?, and (3) do students who use the enhanced intervention have an easier time solving Andes problems than students who use the current version of Andes without dialogue?

Students in high school physics classes taught by four participating teachers, in three Pittsburgh-area schools, will be randomly assigned to one of the two conditions. We will analyze the aggregated data (across schools and classes), and look for interaction effects. Two types of data will be collected: students' test scores—both from their physics course, and from tests that we administer; and system logs. Pre- and posttests will be designed to measure conceptual understanding of physics, qualitative and quantitative problem-solving ability. Since one of the main limitations of previous versions of our reflective dialogue system is that students did not show significant learning gains on a standard measure of conceptual understanding that is similar to the Force Concept Inventory (Hestenes et al. 1992), we will use a similar measure, adapted to the problems covered in the prototype system. At this stage of research, we will simply compare the average gain scores of each condition overall (i.e., on qualitative and quantitative questions), and with respect to questions that address different types of knowledge.

In order to address the third question, regarding differences in students' level of difficulty in solving various types of problems, we will examine the log files that Andes records as students are working on a problem. Specific measures of "difficulty level" obtainable from these logs include the following: counts of how many times students asked for help while solving each problem, how many errors they made, and how many steps they took to solve each problem. We are especially interested in using these measures to compare students across conditions with respect to rotational motion problems since this topic can provide some indication of the intervention's potential to support transfer (in this case, from linear motion to rotational motion).

Findings / Results:

Description of the main findings with specific details.

Since this study is in an early stage, we do not yet have findings and conclusions to report. The proposed poster will describe the project goals, illustrate Rimac with sample reflective dialogues, and describe the "decision rules" guiding automated abstraction and specification of student input that we derive from human tutorial dialogues, and teachers' feedback on the system.

Conclusions:

Description of conclusions, recommendations, and limitations based on findings.

> Not applicable (yet).

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Bloom, B.S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*, 4-16.

Cohen, P.A., Kulik, J.A., & Kulik, C-L.C. (1982). Educational outcomes of tutoring: A Meta-analysis of findings. *American Educational Research Journal 19*(2), 237-248.

Connelly, J., & Katz, S. (2009). Towards more robust learning of physics via reflective dialogue extensions. *Proceedings of Educational Media and Hypermedia.*

Corbett, A. (2001). Cognitive computer tutors: Solving the two-sigma problem. *Lecture Notes in Computer Science: User Modeling 2001*. Volume 2109/2001, 137-147. Springer-Verlag.

Corbett, A., Anderson, J., Graesser, A., Koedinger, K., & VanLehn, K. (1999). Third generation computer tutors: Learn from or ignore human tutors? In *Proceedings of the 1999 Conference on Computer-human Interaction,* 85-86. New York: Association of Computing Machinery.

Craig, S. D., Driscoll, D. M., & Gholson, B. (2006). Constructing knowledge from dialog in an intelligent tutoring system: Interactive learning, vicarious learning and pedagogical agents. *Journal of Educational Multimedia and Hypermedia, 13,* 163–183.

Glod, M. (2007). U.S. teens trail peers around world on math-science test. *Washington Post*. Available: http://www.washingtonpost.com/wp-syn/content/article/2007/12/04/AR2007120400730.html.

Hamblin, A. (2009). President Obama makes commitment to increase research and development funding. *Organizing for America*. Available: http://my.barackobama.com/page/community/post/amyhamblin/gGxW3n.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher, 30*(3):141-158.

Katz, S., Allbritton, D., & Connelly, J. (2003). “Going beyond the problem given: How human tutors use post-solution discussions to support transfer,” *International Journal of Artificial Intelligence and Education,* Volume 13, 1, 79-116.

Katz, S., Connelly, J., & Wilson, C. L. (2005). When should dialogues in a scaffolded learning environment take place? *Proceedings of EdMedia 2005*.

Michaels, S., O'Connor, C., & Resnick, L. (2008). Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. *Studies in Philosophy and Education. 27*(4), 283-297. Springer Netherlands.

Pickering, M.J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral Brain Sciences,* 169-226.

VanLehn, K., Lynch, C., Schulze, K. Shapiro, J. A., Shelby, R., Taylor, L., Treacy, D., Weinstein, A., & Wintersgill, M. (2005). The Andes physics tutoring system: Lessons Learned. In *International Journal of Artificial Intelligence and Education*, 15 (3), 1-47.

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rose, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science* 31(1), 3-62.

Ward, A., Litman, D., (2007). Dialog convergence and learning. (2007). *Proceedings 13th International Conference on Artificial Intelligence in Education*, Los Angeles, CA.

Ward, A., & Litman, D. (2008). Semantic cohesion and learning. In *Proceedings 9th International Conference on Intelligent Tutoring Systems* (ITS), 459-469. Ann Arbor, Michigan.

Ward, A., Connelly, J., Katz, S., Litman, D., & Wilson, C. (2009). Cohesion, semantics, and learning in reflective dialogue. In *Proceedings of the Workshop on Natural Language Processing in Support of Learning: Metrics, Feedback, & Connectivity*, held with the 14th International Conference on Artificial Intelligence in Education, AIED 2009.

Appendix B. Tables and Figures

Not included in page count.

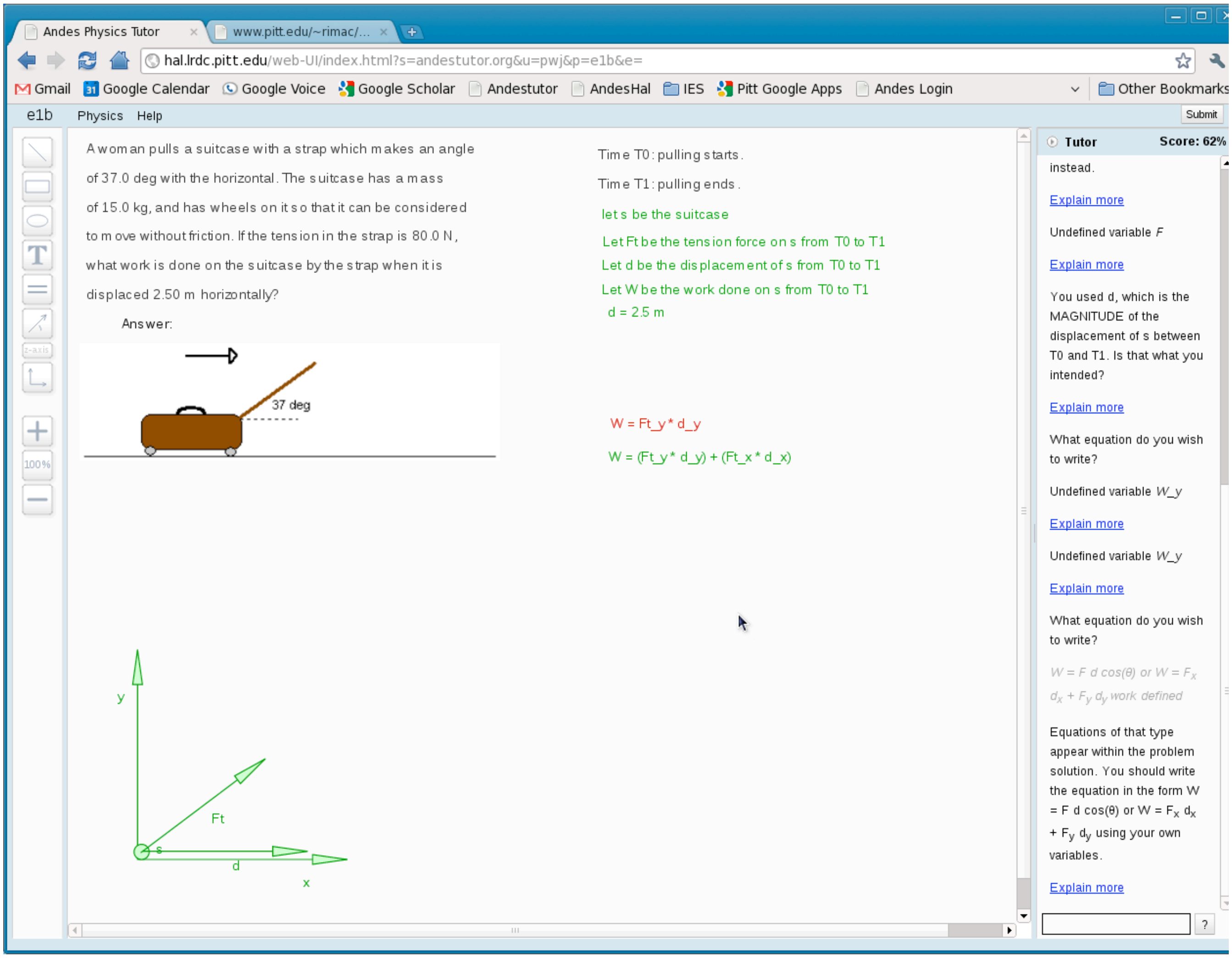

Figure 1: Andes user interface

Andes Problem: A woman pulls a suitcase with a strap which makes an angle of 37.0 deg with the horizontal. The suitcase has a mass of 15.0 kg, and has wheels on it so that it can be considered to move without friction. If the tension in the strap is 80.0 N, what work is done on the suitcase by the strap when it is displaced 2.50 m horizontally?

Reflection Question: Suppose that the suitcase has no wheels, and is instead subject to a frictional force as it slides across the floor. If everything else were the same, how would the work done by the strap on the suitcase change?

Reflective Dialogue

1. **Tutor-1:** What quantities do you need to calculate work?
2. **Student-1:** force, displacement
3. **Tutor-2:** Unless you know both the force and displacement **vectors**, you also need the **angle** between the force and the displacement to calculate work.
4. **Tutor-3:** In short, to calculate work you need the force vector and the displacement vector or, alternatively, the magnitude of the force, the magnitude of the displacement, and the angle between the force and the displacement. What type of force does the strap exert on the suitcase?
5. **Student-2:** tension
6. **Tutor-4:** Good. Does the tension force have a component in the direction of the suitcase's displacement?...

Figure 2. Excerpt of a Sample Reflective Dialogue in Andes