# Pilot Study of a Tutorial Dialogue System that Emulates the Contingent Scaffolding of Human Tutors

Sandra Katz[1], Patricia Albacete[1], Pamela Jordan[1]
Irene-Angelica Chounta[2], Bruce M. McLaren[3]

Learning Research and Development Center, University of Pittsburgh, USA[1]; Institute of Education, University of Tartu, Estonia[2]; Human-Computer Interaction Institute, Carnegie Mellon University, USA[3]
Contact: [katz | palbacet | pjordan]@pitt.edu

**Abstract.** This paper describes an initial classroom evaluation of Rimac, a natural-language tutoring system for physics. Rimac uses a student model to guide decisions about *what content to discuss* next during reflective dialogues that the student and automated tutor engage in after students solve quantitative physics problems, and *how much support to provide* during these discussions—that is, domain contingent scaffolding and instructional contingent scaffolding, respectively (e.g., Wood, 2001). The pilot study compared learning gains of high school students who were randomly assigned to use an experimental version of Rimac, which uses students' responses to pretest items to initialize the student model and dynamically updates the model based on students' responses to the automated tutor's questions during reflective dialogues (Chounta et al., 2017) with a control version of Rimac, which initializes its student model based on students' pretest performance but does not update the model further. Although preliminary results do not reveal a significant difference in learning between conditions, analyses of selected knowledge components indicate that the experimental condition outperforms the control when it has sufficient evidence upon which to update its student model.

## 1. Introduction

Like scaffolds at construction sites, scaffolding by a parent, tutor, or classroom teacher provides the learner with temporary support (Wood, Bruner, & Ross, 1976). However, instructional scaffolding has several key features that distinguish it from its metaphoric source. As described by van de Pol, Volman, and Beishuizen (2010), these features include *contingency*, *fading* and, correspondingly, the gradual *transfer of responsibility* for learning and successful performance to the learner. "Contingency" refers to the adaptive nature of scaffolding and is believed to be its core feature, from which the other two features stem. Instructors dynamically adjust their degree of control over the learning task according to their diagnosis of the student's current level of understanding or performance (Wood & Middleton, 1975). "Fading" refers to the gradual release of this support so that scaffolding can achieve its ultimate aim: to shift responsibility for successful performance to the student.

Wood and Wood (1996) distinguished between three types of contingency during human tutoring sessions: *temporal*, *domain*, and *instructional contingency* (see also Wood, 2003). Temporal contingency is concerned with deciding *when* to intervene versus letting the learner struggle for a while or to request help. Domain contingency is concerned with choosing appropriate content to address during an intervention, while instructional contingency is concerned with deciding how to address focal content—for example, in how much detail and through which pedagogical strategies (e.g., modeling, hinting, explaining, question asking)?

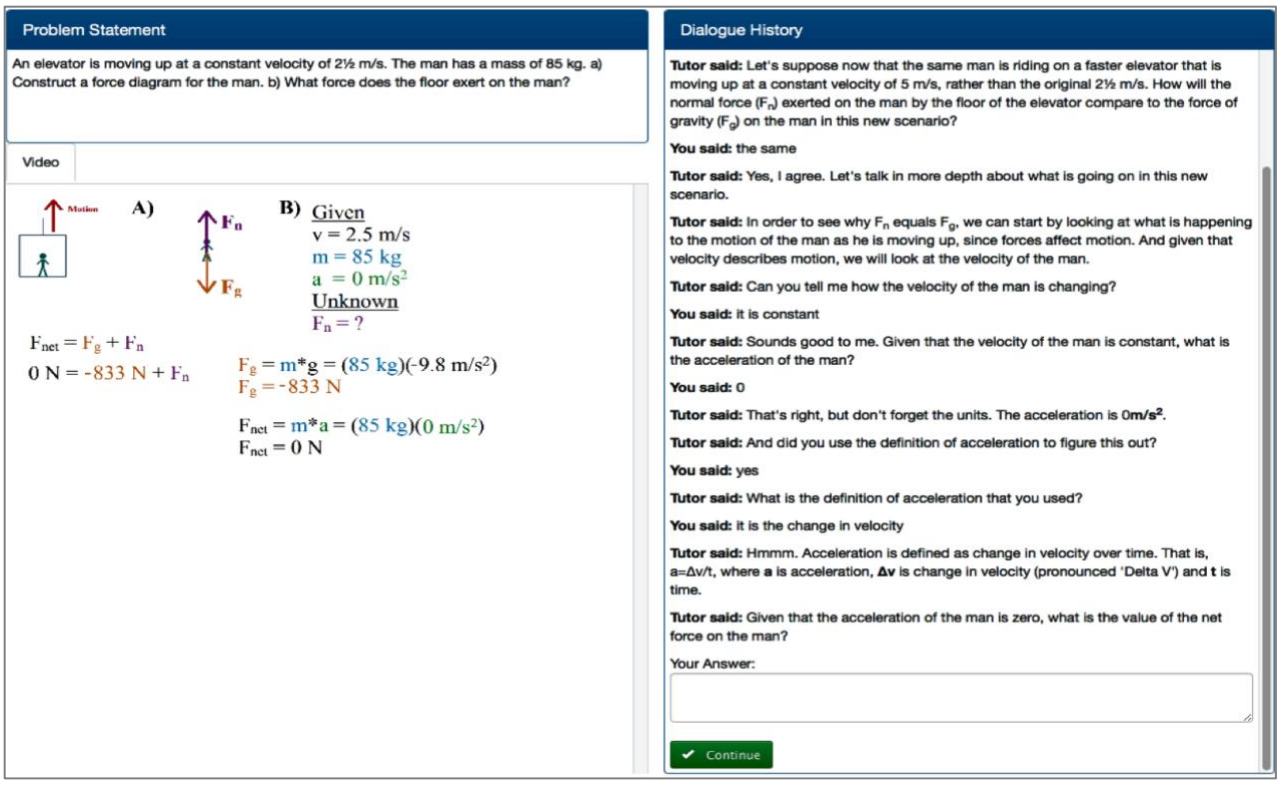


**Figure 1:** Rimac tutor interface. Problem statement shown in upper left pane, worked example video in lower left pane, and dialogue excerpt in right pane.

We have been developing a natural-language tutoring system for physics called Rimac, which emulates these three forms of contingent scaffolding (e.g., Albacete et al., 2018; Katz et al., 2018). The tutor engages students in reflective dialogues after they have solved a physics problem on paper and have then watched a video that presents an annotated "walk-through" of a sample correct solution. (See Figure 1.) These dialogues are developed using an authoring framework called *Knowledge Construction Dialogues* (KDCs) (e.g., Evens & Michael, 2006; Graesser et al., 2004; Rosé et al., 2001). KCDs present a series of carefully ordered questions known as a *Directed Line of Reasoning* (DLR) (Evens & Michael, 2006), which guide students in responding to complex conceptual questions (reflection questions, or RQs). Rimac's student modeling engine informs decisions about what content to address next during a dialogue and how to discuss this focal content. These decisions depend on the student model's assessment of a student's understanding of the knowledge components (KCs)—concepts, principles, etc.—associated with each step of a DLR (Albacete et al., 2018; Katz et al., 2018).

Like other tutoring systems that engage students in KCDs, Rimac implements temporal contingency by intervening when the student responds to the automated tutor's questions incorrectly. The tutor launches a remedial dialogue to address a misconception or knowledge gap indicated by the student's response. Domain contingency in Rimac focuses on choosing the next step in a dialogue's line of reasoning—for example, should the tutor ask the question at the next step in the main line of reasoning or address necessary background content first, if the student model indicates that the student will not be able to respond to the next question in the DLR correctly? Albacete et al. (2018) describe our approach to implementing domain contingency in Rimac; Katz et al. (2018) describe our approach to emulating the instructional contingency of

human tutoring. This paper presents preliminary results of a pilot study to evaluate the tutoring system in high school physics classrooms.

## 2. Methods

### 2.1. Study Design and Hypothesis

The pilot study compared learning gains of high school physics students who were randomly assigned to use one of two versions of Rimac: an experimental version that implements a dynamic approach to contingent scaffolding by continuously updating its student model based on the tutor's assessment of students' dialogue responses, as described in Albacete et al. (2018), and a control version that initializes the student model based on students' pretest scores but does not update this model further. The control group in the current study used a modified version of the system described in Jordan, Albacete, and Katz (2017).[1] We hypothesized that higher learning gains would be realized by students who used the experimental version of Rimac than by students who used the control version.

### 2.2. Participants

191 students from three Pittsburgh-area high schools participated in the study. However, we analyzed the results of one school separately from the two other schools due to differences in the set of problems used across schools. Two schools used the same set of four problems and their associated reflective dialogues; one school used three of these problems but replaced one problem with another due to differences in curricula across school districts. Data from the first two schools were therefore analyzed separately from that of the third school; results will be presented accordingly.

### 2.3. Procedure

A pretest and isomorphic posttest, consisting of 36 multiple choice questions were developed. Participants took the pretest in class. After the pretest, students interleaved solving homework problems on paper with using the system as follows: First, students solved a problem on paper. They then viewed a video of the solution to the problem on the system, which contained no reference to any conceptual aspect of the problem. They then engaged in several reflective dialogues with the automated tutor (2-4 dialogues per problem); the dialogues addressed conceptual aspects of the recently solved problem. After all problems and dialogues were completed, students took an in-class posttest. Thirty minutes was allowed for completion of each test.

[1] As in the version of Rimac described in Jordan et al. (2017), the control version of Rimac in the current study allowed a student to skip a reflection question (RQ) if the student answered the RQ correctly and the student's scores on pretest items that addressed the most important knowledge components associated with that RQ were at or above threshold (80%). If the student's scores on these prerequisite KC's were below this threshold, the tutor addressed relevant background knowledge at a level deemed appropriate, given the student's pretest scores on these highly relevant KC's. The version of Rimac used in the current study differs from the previous version (Jordan et al., 2017) in how it handles incorrect responses to an RQ. Specifically, instead of adjusting the amount of background knowledge addressed at each step of the dialogue's line of reasoning according to students' pretest performance, the tutor takes a more cautious approach, addressing all prerequisite knowledge at each step.

## 3. Results

**Schools 1 and 2.** One hundred and eighteen students taking high school physics participated in the pilot study across these two schools. They were randomly assigned to condition (experimental = 60; control = 58). Students were assigned to complete four of the five problems and associated dialogues in the pilot study version of Rimac as homework. (Problems are referred to as "elevator", "football", "sled", and "arrow", reflecting the main physical objects that they refer to.) Data analysis focused on 29 of the 36 test items.[2]

There was no difference in incoming knowledge between conditions, as measured by the pretest ($t(116)=.989$, p=.325). Overall, students in these two schools learned from interacting with the tutoring system, as indicated by a statistically significant difference between mean pretest and posttest scores: $t(117)=5.203$, $p<.001$ considering all students; $t(59)=4.020$, $p<.001$considering the experimental group, and $t(57)=3.350$, $p=.001$considering the control group.

Although both conditions showed learning gains, there was no statistically significant difference in gains across conditions. An ANCOVA controlling for the effect of prior knowledge (as measured by pretest) revealed a non-significant effect of condition on pretest to posttest gain score ($F(1,115)=.008$, $p=.927$). Neither an aptitude treatment interaction ($F(1,114)=0.000$, p=.997)) nor a difference in time on task ($t(116)=.342$, $p=.733$) was observed between conditions.

**School 3.** Seventy-three students taking high school physics participated and were randomly assigned to conditions (experimental = 31; control = 42). Students in this school solved three of the same four problems as did students in Schools 1 and 2, but did not solve the "sled" problem; they solved the "amusement park rider" problem instead. Data analysis included 35 of the 36 test items, eliminating one item that pertained to the "sled" problem.

As with the other schools, students in School 3 learned from interacting with the system, as measured by a statistically significant difference between mean pretest and posttest scores: $t(30)=6.540$, $p<.001$, considering the experimental group; $t(41)=7.565$, $p<.001$, considering the control group. There was no difference in incoming knowledge between conditions, as measured by the pretest ($t(71)= -.127$, $p=.900$), and there was not a significant difference in learning gains between conditions ($t(71)= 1.226$, $p=.224$). Similarly, an ANCOVA measuring the effect of condition on posttest controlling for the effect of pretest was insignificant ($F(1,70) = 1.770$, $p=.188$).

## 3. Discussion

We suspect that the lack of a difference in learning gains between conditions might be due to limited exposure to knowledge components (KC) during the dialogues. Conversely, we predict that more frequent exposure to KCs will result in greater learning of those KCs than less frequent exposure. As an initial test of this interpretation, we examined the relation between student performance in terms of pretest to posttest gain score and frequency of KC references (exposure). Towards this end, we categorized KCs addressed in the dialogues based on how frequently they are referenced in the tutorial dialogues. Specifically, we defined three KC groups:

1. KCs with no or one reference (reference <= 1)
2. KCs with an average number of references (1 < references < 5)

[2] Seven test items were eliminated because they addressed a problem ("amusement park rider") that students in these two schools did not work on, since it applied physics concepts that were not covered by the schools' physics curricula.

3. KCs with many references (references >= 5)

To clarify, "references" means the number of times a student came across a KC during the course of the reflective dialogues, either in the tutor's feedback or questions.

For each student, we calculated the number of references per KC and the Learning Gain (LG) for each KC as the difference between the posttest and the pretest KC score. Then, we computed the average learning gain per KC for each of the three KC groups. Table 1 presents the results for the three KC groups. For example, the last row of Table 1 shows that on average 11 KCs were referenced 5 or more times. A KC that belongs to the "frequently" referenced KCs group was on average practiced 6.4 times. The learning gain on average for these frequently referenced KCs is 1.01 from pretest to post-test. The learning gain difference between the "no or one reference" group and the "average/many references" groups is statistically significant ($p<0.01$). This result provides preliminary support for our interpretation of the results of the pilot study: that learning gains are dependent on KC exposure. It is possible that students did not get enough practice with KCs in the dialogues for the dynamic updating of the student model in the experimental condition to make a difference in learning relative to the control condition, with its fixed student model.

With respect to the student clusters (high, medium, low incoming knowledge), the analysis suggested that the high prior knowledge students got fewer practice opportunities (or references) per KC than the medium and low prior knowledge students. For example, a frequently occurring KC would be referenced 6.6 times on average for a high prior knowledge student and 11 times on average for a medium or low prior knowledge student. This difference is statistically significant ($p<0.05$) and expected from a contingency standpoint. Since high incoming knowledge students will make fewer mistakes, the adaptive tutor will discuss concepts less frequently than with low incoming knowledge students.

**Table 1.** Results for frequently and infrequently practiced KCs[3]

| KC Groups based on frequency of reference | Average number of KCs in the respective group | Average number of references per KC in the respective group | Average knowledge gain per KC in the respective group |
|---|---|---|---|
| **0 or 1** | 64 (13) | 0.29 (0.06) | 0.029 (0.09) |
| **Average (2-5)** | 43 (9) | 2.65 (0.13) | 0.073 (0.12) |
| **Many (5 or more)** | 11 (7) | 6.4 (0.87) | 1.01 (0.2) |

## 4. Conclusion

If verified through further classroom trials, the results of the pilot study described in this paper indicate that a natural-language tutoring system that uses a student model to drive contingent scaffolding during conceptual, reflective dialogues has the potential to support learning better than

[3] Standard deviations are in parentheses, following each reported average.

does a system with a pretest-initialized student model. However, these differences will only be observed if the student model has sufficient evidence of students' grasp of particular KCs—that is, KC exposures or references. In future work, we plan to allow students to inspect their student model and negotiate changes to it, by implementing an Open Learner Model (OLM) (e.g., Bull & Kay, 2007). We expect that when students try to persuade the system to increase (or, less likely, decrease) its assessment of the student's understanding of particular KCs, and the tutor issues a question to test the student's self-assessment, students' responses will boost the evidence base needed to improve student model accuracy.

## References


Albacete, P., Jordan, P., Lusetich, D., Chounta, I., Katz, S., & McLaren, B. M. (2018). *Providing Proactive Scaffolding During Tutorial Dialogue Using Guidance from Student Model Predictions*. Paper presented at the The 19th International Conference on Artificial Intelligence in Education, London, UK.

Bull, S., & Kay, J. (2007). Student models that invite the learner in: The SMILI:() Open learner modelling framework. *International Journal of Artificial Intelligence in Education, 17*(2), 89-120.

Evens, M., & Michael, J. (2006). *One-on-one tutoring by humans and computers*: Psychology Press.

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A., & Louwerse, M. M. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods, 36*(2), 180-192.

Jordan, P., Albacete, P., & Katz, S. (2017). *Adapting Step Granularity in Tutorial Dialogue Based on Pretest Scores.* Paper presented at the International Conference on Artificial Intelligence in Education.

Katz, S., Albacete, P., Jordan, P., Lusetich, D., Chounta, I.-A., & McLaren, B. M. (2018). Operationalizing contingent tutoring in a natural-language dialogue system. In S. Craig (Ed.), *Tutoring and Intelligent Tutoring Systems*. New York: Nova Science Publishers.

Rosé, C. P., Jordan, P., Ringenberg, M., Siler, S., VanLehn, K., & Weinstein, A. (2001). *Interactive conceptual tutoring in Atlas-Andes.* Paper presented at the Proceedings of AI in Education 2001 Conference.

van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher–student interaction: A decade of research. *Educational Psychology Review, 22*(3), 271-296.

Wood, D. (2001). Scaffolding, contingent tutoring, and computer-supported learning. *International Journal of Artificial Intelligence in Education, 12*(3), 280-293.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of child psychology and psychiatry, 17*(2), 89-100.

Wood, D., & Middleton, D. (1975). A study of assisted problem-solving. *British Journal of Psychology, 66*(2), 181-191.