Exploring the Effects of Redundancy within a Tutorial Dialogue System: Restating Students' Responses

Pamela Jordan **Patricia Albacete** **Sandra Katz**
Learning Research and Development Center
University of Pittsburgh
pjordan@pitt.edu

Abstract

Although restating part of a student's correct response correlates with learning and various types of restatements have been incorporated into tutorial dialogue systems, this tactic has not been tested in isolation to determine if it causally contributes to learning. When we explored the effect of tutor restatements that support inference on student learning, it did not benefit all students equally. We found that students with lower incoming knowledge tend to benefit more from an increased level of these types of restatement while students with higher incoming knowledge tend to benefit more from a decreased level of such restatements. This finding has implications for tutorial dialogue system design since an inappropriate use of restatements could dampen learning.

1 Introduction

A tutor restating part of a student's dialogue contribution can be motivated by a range of communicative intentions (e.g. a tutor intends to reformulate a response, so that it is correct) and at the surface level can range from exact repetitions, to using different words while keeping the content semantically equivalent, to semantic reformulations which are often prefaced by markers such as "in other words" and "this means that" (Hyland, 2007). Some of the intentions associated with reformulations in the context of classroom lectures (Murillo, 2008) that also appear in human tutorial dialogue (Jordan et al., 2012) include, among others, definition (reformulate a prior statement so terms are defined), correction (reformulate a prior statement so it is correct) and consequence (reformulate so implications of a prior statement are clear).

But restatements also have intentions unique to the context of interactive discourse. We observed that human tutors, like classroom teachers who encourage and support discussion, frequently implement two types of restatement moves: revoicing and marking. Revoicing is characterized by a reformulation of what the student said. Like classroom teachers who facilitate discussions using a technique called "Accountable Talk" (O'Connor and Michaels, 1993), tutors sometimes revoice in order to verify their understanding of what a student was trying to say and, in the case of a correct student contribution, perhaps to model a better way of saying it. Marking, on the other hand, emphasizes what the teacher or tutor considers most important in what the student said and attempts to direct the student to focus his/her continued discussion on that.

Several recent studies of human tutorial dialogue have looked at particular aspects of restatements, for example, (Chi and Roy, 2010; Becker et al., 2011; Dzikovska et al., 2008; Litman and Forbes-Riley, 2006). One study examines face-to-face naturalistic tutorial dialogue in which a tutor helps a student work through a physics problem (Chi and Roy, 2010). The authors suggest that when the tutor repeats part of what the student said, it is often done with the intention of providing positive feedback for correct answers. Another of these recent studies collected a corpus using trained human tutors who filled in for a conversational virtual tutor in a science education system (Becker et al., 2011) and noted that a restatement can help a student who is struggling with a particular concept by modeling a good answer and can mark an aspect of the student's response to focus on in the ongoing discussion. Below we show excerpts from our corpus of human-human typed dialogues that illustrate these uses of restatement.

> T: How do we know if there is a net force on the bullet in this problem?

Proceedings of the SIGDIAL 2015 Conference, pages 51–59,
Prague, Czech Republic, 2-4 September 2015.

S: *if m*a does not equal 0*
T: Right, **if the bullet is accelerating it must have a net force on it** - [tutor restatement to mark and provide positive feedback]
T: how do we know it is accelerating?

T: What is speed?
S: *it is velocity without direction*
T: **Right, The (instantaneous) speed is the magnitude of the (instantenous) velocity.** [tutor restatement to model a good answer and provide positive feedback]

Because restatements of correct responses have been shown to correlate with learning (Dzikovska et al., 2008), this suggests the possibility that restatements could causally contribute to learning. While restatements of various types have been incorporated into a number of tutorial dialogue systems, restatement has not been tested in isolation from other tactics to determine whether it has any causal connection to learning. Examples of tutorial dialogue systems that have incorporated restatement include: AutoTutor (Person et al., 2003) where elaborations and summaries often include restatements, CIRCSIM-Tutor (Freedman, 2000), which restates students' answers that are nearly correct except for terminology, and Beetle II (Dzikovska et al., 2008), which restates the correct parts of students' nearly correct or partially correct answers.

Here, we explore the effects on student learning of a tutor's restatement of the student's correct response in the context of a consequence intention (Murillo, 2008)–that is, making an inference explicit as shown in the excerpt below from our corpus.

T: How do we know that we have an acceleration in this problem?
S: because velocity starts at zero, and since the stone is falling, it doesn't remain at zero, thus there is *a change in the velocity* of the stone
T: Ok so because there is **a change in velocity** then there has to be an acc [sic] right? [tutor restatement of correct response while making its implications clear]

We test two alternative hypotheses about this type of restatement: 1) that it will benefit students and 2) that its effect varies according to students' incoming knowledge.

Our discussion of the study that we conducted to test our hypotheses will proceed as follows. First we discuss the motivation for our hypotheses and then we describe the existing tutorial dialogue system we used as a platform for conducting our experiments with three different populations of students. We characterize the degree of restatement supported by the unaltered system and the modifications we made to produce a high restatement and a low restatement version of the system. Next we describe the experimental design and discuss our results in relation to two earlier experiments using different populations and test materials. We conclude by summarizing our results and plans for future work.

2 Background

From the perspective of memory encoding, storage and retrieval (McLeod, 2007), simply repeating back a student's correct answer may have an effect similar to maintenance rehearsal which would just maintain it in the student's working memory but do little to aid transfer to long-term memory. However, connecting the correct answer to something else, which a consequence restatement would do, may have more of an elaborative rehearsal effect which is better for transfer to long-term memory (McLeod, 2007). But the effect may not be applicable for very low incoming knowledge students who are not correct often. Conversely, if the correct answer is already more strongly established in the student's long-term memory–as may be the case for high incoming knowledge students–then restating it could be detrimental, whether the tutor's restatement only acknowledges the student's correct answer or is in the context of a consequence. In this situation it may be better to focus on strengthening the connection between the correct knowledge and other knowledge by having the student recall the correct knowledge on his/her own when it is needed.

From the perspective of interactions between communication strategies and cognitive processing, simulations with artificial agents showed that task performance varied as communication strategies and cognitive processing limits varied (Walker, 1996; Jordan and Walker, 1996). For example, under certain conditions as attention became more limited, repetition of mutually known information displaced from attention other critical problem-solving knowledge for the "hearer" while, conversely, such redundancies could become beneficial when attention was less limited. Possibly a student should not have mutually known information repeated when they are deep in thought (i.e. the processing load is high), because it could displace critical knowledge. On the

other hand, a student who may be having trouble getting started on a question (i.e. the processing load may be lower), may find the repetition beneficial because there is less chance of displacement. The former case may more often describe a high-knowledge student and the latter a low-knowledge student.

Two other strands of research in psychology that are related to our hypotheses examined the effect of text cohesiveness on comprehension for low-knowledge and high-knowledge readers. The first found that unpacking the inferences in text supports comprehension among low-knowledge readers, while less cohesive (higher inference-inducing) text is better suited for high-knowledge readers (McNamara et al., 1996). Forcing the student to figure out what led to a consequence when no premise is explicitly provided could make it similar to a higher inference-inducing text. Reduced cognitive load is a proposed alternative explanation for the "cohesion reversal effect", particularly for high-knowledge readers, who must reconcile their existing schema about the topic discussed in the text with the background material provided in a "highly coherent" text (Kalyuga and Ayres, 2003). High-knowledge students might benefit more from less frequent consequence restatements because these students can make more inferences on their own. Frequent consequence restatements might entail more frequent schema alignment, and therefore an increased cognitive load. However, both of these explanations of the cohesion reversal effect, with respect to high knowledge students (prompted inference-making, or increased cognitive load), may be less plausible for consequence restatement during tutorial dialogue than for reading, because the former involves a proposition that was recently explicitly covered in the dialogue.

3 Experimental Platform

We used an existing natural-language tutoring system, Rimac, to conduct our experiments. It is a web-based system that aims to improve students' conceptual understanding of physics through typed reflective dialogues (Katz and Albacete, 2013). Rimac was built using the TuTalk natural language (NL) tutorial dialogue toolkit (Jordan et al., 2007). Thus its dialogue can be represented as a finite state machine where each state represents a tutor turn. The arcs leaving a state correspond to all classifications of a student's response to the tutor's turn. When a student turn is received, the system determines which arc it best represents and this in turn indicates what tutor state to transition to next. In the context of restatements, because the arc that is the best classification of the student's response leads to a particular tutor state, the tutor state can include that arc in its representation and can easily restate that arc. Note that this simplified approach will produce more reformulations than exact repetitions of student responses but both are acceptable for our experiment.

For this experiment we used Rimac's dynamics content which covers three problems with two reflection questions per problem. These problems and their associated reflective dialogues (two dialogues per problem) were developed in consultation with high school physics teachers. The reflection question dialogues are tutor-initiative only. The tutor primarily asks short answer questions, to keep accuracy of automatic recognition of student responses high. However, the dialogues include some questions that prompt the student for explanations at key points and then presents a menu of responses to which students are directed to map their previously typed explanation. We expect there to be a comparable frequency of misclassification of student responses across the two versions of the system that we created for our experiments since we made no modifications to any student response arcs in the original system.

To create a high restatement system for this experiment, three dialogue content authors independently reviewed and cross-reviewed all of the tutor states in the dialogue specifications for the base system and added tutor restatements of student responses that occurred in two dialogue contexts. Those contexts were: 1) an explicit if-then statement or question that draws the "if" or "then" part from the student's immediately preceding response and 2) a statement or question at the end of a sub-dialogue that draws upon the student's responses during the sub-dialogue. An example for contexts 1 and 2 are shown below for the high restatement condition:

> Context 1:
>
> T: Now let's look at our case in which we have a flying arrow. While the arrow is flying is anything touching or in contact with it?
> S: No [there is nothing touching the arrow during its flight]

T: I agree. Hence since **there is nothing touching the arrow during its flight** there is no contact force applied to it.

Context 2:

T: What is the initial velocity of the faster arrow in the vertical direction as it leaves the bow?
S: [the initial velocity is] 0
T: I agree. Now does the faster arrow have an acceleration in the vertical direction when it leaves the bow?
S: yes, [it has a vertical acceleration of] 9.8 m/s^2
T: Okay! Now that we know **the initial vertical velocity of the fast arrow is zero and that it has a vertical acceleration of 9.8 m/s^2**, what will the value of vertical velocity of the arrow be a second after it leaves the bow?

To create the low restatement system, the same dialogue content authors reviewed the dialogue specifications for all instances of consequence restatements and either deleted or replaced the restatement of the student's response with an anaphoric expression. But the content authors were further directed to not carry out any modification that would disturb the coherency of the dialogue. The low restatement version of the first example above would be identical, except for the second tutor turn, which would read: "I agree. Hence there is no contact force applied to it." The example for context 2 is shown below for the low restatement condition.

Context 2:

T: What is the initial velocity of the faster arrow in the vertical direction as it leaves the bow?
S: 0
T: I agree. Now does the faster arrow have an acceleration in the vertical direction when it leaves the bow?
S: yes, 9.8 m/s^2
T: Okay! Now given what we know about the fast arrow, what will the value of vertical velocity of the arrow be a second after it leaves the bow?

After the experiments (described below) were completed, one of the authors of this paper reviewed the tutor states in the base system and the high and low restatement systems to characterize the number of changes made to create the high and low restatement systems from the base system. These findings are shown in Table 1 in the columns "possible". The "other" restatements, as shown in column 3 of Table 1, include restating the correct part of a partially correct answer and restating a correct answer when it required deeper reasoning to produce. These remain because they were deemed essential to tutoring. Ideally the number of "other" restatements should be equal for "high" and "low". Content authors were instructed to remove repetitions of fully correct answers to simple short answer questions but some were missed for "high". In addition, some restatements that were added to increase consequence for "high" were instead simple repetitions. However, we do not expect simple repetitions to affect learning, especially when their frequency is low, as reflected in the "avg" columns.

Table 1: Modifications to create the high and low restatement systems from the base system (labeled "possible") and the average number of states students experienced (labeled "avg")

System	**Number of Restatement States**			
	Consequence		**Other**	
	possible	avg	possible	avg
Base	48	NA	18	NA
High	77	19.8	19	2.6
Low	4	.8	7	.375

4 Methods

Participants Our comparison of the high and low restatement versions of Rimac was conducted during high school physics classes at three schools in the Pittsburgh PA area. The study followed the course unit on dynamics with a total of 168 students participating. Students were randomly assigned to one of two conditions: high restatement (N= 88; 30 females, 58 males) and low restatement (N= 80; 27 females, 53 males).

Materials Students interacted with either a high or low restatement version of Rimac, as described in the previous section, to discuss the physics conceptual knowledge associated with three quantitative dynamics problems.

We developed a 21 item pretest and isomorphic post-test (that is, each question was equivalent to a pretest question, but with a different cover story) to measure learning differences from interactions with the system. The test included nine multiple choice problems and twelve open response problems and focused on testing students' conceptual understanding of physics instead of their ability to solve quantitative problems.

Procedure On the first day, the teacher gave the pretest in class and assigned the three dynamics problems for homework. During the next one to two class days (depending on whether classes

Table 2: Learning from interacting with the systems, for both conditions combined and separately for the high and low restatement conditions

Problems	**Condition**	**Pretest Mean (*SD*)**	**Posttest Mean (*SD*)**	***t*(*n*), *p***
All	Combined	7.90 (2.40) 0.376 (0.114)	8.97 (2.88) 0.427 (0.137)	$t(167)=5.60$, $p<0.01$
	High	7.71 (2.36) 0.367 (0.113)	8.73 (2.73) 0.416 (0.130)	$t(87)=3.56$, $p<0.01$
	Low	8.11 (2.44) 0.386 (0.116)	9.23 (3.02) 0.440 (0.144)	$t(79)=4.49$, $p<0.01$
Multiple-choice	Combined	4.73 (1.40) 0.525 (0.156)	5.20 (1.50) 0.578 (0.167)	$t(167)=3.63$, $p<0.01$
	High	4.67 (1.37) 0.519 (0.152)	5.16 (1.46) 0.573 (0.162)	$t(87)=2.73$, $p=0.01$
	Low	4.79 (1.44) 0.532 (0.160)	5.25 (1.55) 0.583 (0.173)	$t(79)=2.39$, $p=0.02$
Open-response	Combined	3.18 (1.48) 0.265 (0.124)	3.77 (1.78) 0.314 (0.148)	$t(167)=5.38$, $p<0.01$
	High	3.04 (1.47) 0.253 (0.123)	3.57 (1.68) 0.298 (0.140)	$t(87)=3.13$, $p<0.01$
	Low	3.32 (1.49) 0.277 (0.124)	3.98 (1.87) 0.332 (0.156)	$t(79)=4.8$, $p<0.01$

were approximately 45 min. or 80 min. long), students watched a video of a sample, worked-out solution to each homework problem in one of the two versions of Rimac and engaged in two "reflective dialogues" after each problem-solving video. The videos demonstrated how to solve the problem only and did not offer any conceptual explanations. Hence we do not believe that the videos contributed to learning gains. Finally, at the next class meeting, teachers gave the post-test.

5 Results

We evaluated the data to determine whether students who interacted with the tutoring system learned, as measured by gain from pretest to post-test, regardless of their treatment condition (i.e. which version of Rimac they were assigned to use), and if there was an aptitude-treatment interaction; in particular, an interaction between students' prior knowledge about physics (as measured by pretest score) and how much students learned in each condition (as measured by gain score).

The data was first analyzed considering all problems together and then multiple-choice and open-response problems were considered separately. The rationale for this further division of test items is that open-response problems, unlike multiple-choice problems, would allow us to determine whether students are able to verbalize coherent conceptual explanations of the physics phenomena tested in these problems. Moreover, open-response problems do not allow for guessing of the correct answer to the extent that multiple-choice test items do.

Learning Performance & Time on Task To determine whether interaction with the system, regardless of condition, promoted learning, we compared pretest scores with post-test scores. Towards this end, we performed paired samples t-tests. When all students were considered together, we found a statistically significant difference between pretest and post-test scores for all problems together, multiple-choice problems, and open-response problems as shown in Table 2. When students in each condition were considered separately, we again found a statistically significant difference between pretest and post-test for all problems together, multiple-choice problems, and open-response problems as shown in Table 2. These results suggest that students in both conditions learned from interacting with the system.

Prior to testing for differences between conditions, we tested for a difference in time on task between conditions. No statistically significant difference was found between conditions for the mean time on task.

High Restatement vs. Low Restatement First, we confirmed that there was significantly more consequence restatement in the high restatement condition than in the low restatement condition using independent samples t-tests: M(high)=20.4,

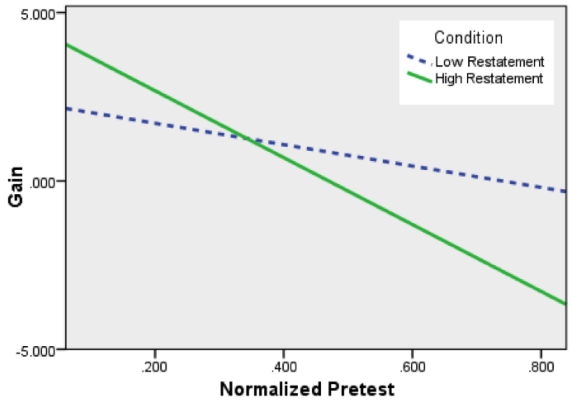

Figure 1: Prior knowledge-treatment interaction for All Problems

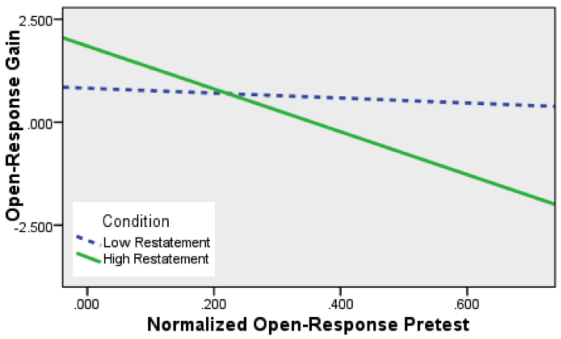

Figure 2: Prior knowledge-treatment interaction for Open-Response Problems

M(low)=.8; $t(91)=29.3, p<.0001$. Next, to test whether students who used the high restatement version of the system would perform differently from students who used the low restatement version, we compared students' gains from pretest to post-test between conditions using independent samples t-tests. Gains were defined as (post-test - pretest) and their normalized versions as (post-test/#problems) - (pretest/#problems).[1]

We found no significant differences in gains between conditions for any subset of problems. This suggests that the presence or absence of a consequence restatement has the same effect on learning when students of all knowledge levels are considered together.

Prior knowledge-treatment interaction To investigate whether there was a prior knowledge treatment interaction, we performed a multiple regression analysis using condition, prior-knowledge (as measured by pretest) and condition * prior-knowledge (interaction) as explanatory variables, and gain as the dependent variable. When all problems were considered together, we found a significant interaction between condition and prior knowledge in their effect on gains (t=-2.126,p=0.04). Likewise, we found a significant interaction when we considered only gains on open-response problems (t=-2.689,p=0.01). However, for multiple-choice problems we did not find a significant interaction.

The graph of gain vs. prior knowledge in Figure 1 shows the fitted lines for both conditions when considering all problems. It suggests that students with pretest scores that are 35% correct (7.5) or less benefit more from the high restatement version of the system than from the low restatement version. However students with pretest scores above 35% correct benefit more from the low restatement version of the system. The graph of gain vs. prior knowledge for open-response problems is shown in Figure 2. It suggests that students with pretest scores of 23% or less on open-response items benefit more from higher restatement and students with pretest scores greater than 23% benefit more from lower restatement. Both findings offer evidence to support the hypothesis that the effect of consequence restatements varies according to students' incoming knowledge. In particular, it suggests that lower knowledge students benefit more from high restatement in inferential contexts while higher knowledge students benefit more from low restatement.

6 Additional Support for a Prior Knowledge-Treatment Interaction from Earlier Experiments

Prior to the study that we described in Section 5, which we will refer to now as experiment E3, we conducted two field trials, E1 and E2, which differed only by the versions of the tests that we administered and the populations recruited. We will refer to the test we previously described in Section 4 as T3, to distinguish it from the tests administered during the prior experiments (T1 and T2).

Field Trial E1 with test T1 The first field trial, E1, utilized undergraduate students only and test T1. We recruited undergraduates (N=62) who had taken only high school physics within the last two

[1]The reason for using both measures is that each measure relates the same information, but in a different way. The full test scores show means and standard deviations in terms of number of problems solved correctly (given that each test item has a score of 0-1) whereas the normalized values convey the same results in terms of percent of correct responses.

years. The goal was to sample students whose knowledge of physics was similar to that of our target high school population. Test T1 was used in previous experiments with high school students for the dynamics domain.

Just as with E3, we confirmed that there was significantly more consequence restatement in the high restatement condition than in the low restatement condition using independent samples t-tests: M(high)=24.2, M(low)=1.2; $t(36)=45.7, p<.0001$. Similarly, we found that for the undergraduate population there were no significant differences in gains between conditions. However, for this population there were no significant interactions between conditions and prior knowledge. Since we had found a prior knowledge treatment interaction in experiment E3, we re-examined the pretest scores of the undergraduates, to investigate whether students' incoming knowledge could have been a factor.

We found that the pretest mean for the undergraduates was 44% correct (SD=14%) while the pretest mean for the high school students who had taken test T1 was lower at 37% correct (SD=13%). Furthermore, the high school students who had taken T1 had a post-test mean of 40% correct (SD=16%) which was lower than the **pretest** mean of E1's undergraduates. The undergraduates' prior knowledge is clearly higher than that of the high school students. Given the higher prior knowledge of the undergraduates in E1 (compared with the high school students who had taken T1), we expected that the mean gain for the low restatement condition in E1 (M=2.71, SD=2.18; normalized M=.12, SD=.10) would tend to be higher than for the high restatement condition (M=1.99, SD=2.24; normalized M=.09, SD=.10) and that was the case.

Hence, this pattern is consistent with the second hypothesis that the effect of consequence restatements varies according to incoming knowledge. While there was no significant difference between conditions for the undergraduate population, undergraduates had higher prior knowledge than high school students and for undergraduates the mean gain for the low restatement condition was higher than for the high restatement condition which is in the same direction as the findings for E3.

Field Trial E2 with test T2 We decided to refine test T1, which was used in E1, to create test T2. We used test T2 in field trial E2 with high school students (N=88) who were from two different local high schools from those who participated in experiment E3.

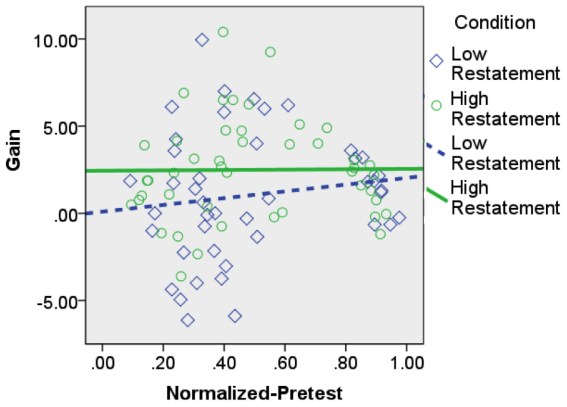

Figure 3: Prior knowledge-treatment interaction for All Problems for E2

As before with E3, we confirmed that there was significantly more consequence restatement in the high restatement condition than in the low restatement condition using independent samples t-tests: M(high)=20.3, M(low)=.64; $t(56)=21.8, p<.0001$. With this population, however, we found statistically significant differences in learning gains between conditions that favored the high restatement version of the system. Using independent samples t-tests, we found significant differences for all test problems together: M(high)=2.49 SD=2.90, M(low)=1.04 SD=3.68; $t(86)=2.07, p<.04$ and for multiple-choice problems: M(high)=.66 SD=1.27, M(low)=-.-6 SD=1.4; $t(86)=2.51, p<.01$ but not for open-response problems. However, there were no statistically significant interactions between condition and prior knowledge for any subset of test problems.

Given the results of experiment E3 and the pattern in E1, we re-examined the pretest scores of these high school students to consider whether their incoming knowledge could have been lower than the students in E3. The graph of the gain vs. pretest scores in Figure 3 shows that gains for students in the high restatement condition were better than for students in the low restatement condition. However, the difference was more pronounced for lower incoming knowledge students than for higher incoming knowledge students which agrees with the pattern in E3. Moreover, one of the schools in this sample had a significantly lower pretest mean than the other school (M=36%, SD=16% vs. M=86%, SD=8%;

t(86)=14.9,p=.000) and a larger sample size (N=65 vs. N=23). This suggests there were more lower incoming knowledge students in E2 than higher incoming knowledge students.

So there is a pattern that is consistent with the finding in E3 and the pattern in E1. The results suggested that the high restatement condition was significantly better than the low restatement one; however, more of the population seemed to have lower incoming knowledge which would favor the high restatement condition. However, more experimentation with populations similar to these two schools is needed. It is possible that the incoming knowledge in this one school is comparable to the ones in E3. This was the only high-school in which we had to move from the classroom to a computer lab. This added disruption to the usual classroom routine may have made it more difficult for students to "settle in" and concentrate. If the students had problems focusing, then the added repetitions may have been helpful.

Experiment E3 with test T3 After E2, we shortened the test to create T3, which was used in experiment E3, the focus of this paper. While the tests differed across all three experiments, so we cannot directly compare the populations, the patterns in each case seem consistent with the prior knowledge treatment interactions that we found in study E3, as reported in Section 5. However, experiments that use the same test would be necessary to verify these patterns.

7 Conclusions and Future Work

We found that students learned from the tutoring system, across conditions, as measured by differences in pre-test and post-test scores. In the main study reported here (E3), there was no difference in learning gains between conditions, which suggests that the presence or absence of consequence restatement in a system has a similar effect for all students considered together; that is, irrespective of their prior knowledge. However, we did find a prior knowledge treatment interaction which supported the hypothesis that the effect of consequence restatement varies according to students' prior knowledge. In particular, our results suggest that lower knowledge students would benefit more from a high restatement system while higher knowledge students would benefit more from a low restatement system.

Two earlier studies with different populations and tests also support this finding. While there was no significant difference in learning gains between conditions for the study with the undergraduate population (E1), undergraduates had higher prior knowledge than high school students and for undergraduates the low restatement condition had a higher mean gain than the high restatement condition. For the earlier study with a different set of high schools (E2), there was a significant difference in learning gains between the high and low restatement conditions that favored the high restatement condition but more of the population seemed to have lower incoming knowledge which would favor that condition. Moreover, the lower the student's incoming knowledge, the larger the benefit of high restatement. However, these results are preliminary and require further experimentation to better understand when and why consequence restatements can support learning.

The findings across the three experiments suggest that system designers may need to be careful in their use of restatement as it may dampen learning if there is a mismatch with students' prior knowledge levels. Further it suggests that when building tutorial dialogue systems, care must be taken in the tactics and strategies that may be applied to address system limitations. For example, spoken dialogue systems sometimes use an explicit confirmation strategy to address repeated speech recognition errors (Litman and Pan, 2000). Carrying such a strategy over to tutorial applications could have an unintended impact on some students' learning outcomes.

In future research, we plan to determine if the benefits of the high and low restatement versions of Rimac can be used advantageously in a system that adapts to students' knowledge levels and to formulate and test additional hypotheses for other types of restatement.

Acknowledgements.

We thank Stefani Allegretti, Michael Lipschultz, Diane Litman, Dennis Lusetich, Svetlana Romanova, and Scott Silliman for their contributions. This research was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A130441 to the University of Pittsburgh. The opinions expressed are those of the authors and do not necessarily represent the views of the Institute or the U.S. Department of Education.

References

L. Becker, W. Ward, S. Van Vuuren, and M. Palmer. 2011. Discuss: A dialogue move taxonomy layered over semantic representations. In *IWCS 2011: The 9th International Conference on Computational Semantics*, Oxford, England, January.

M. T. H. Chi and M. Roy. 2010. How adaptive is an expert human tutor? In *10th International Conference on Intelligent Tutoring Systems (ITS)*, pages 401–412.

M. Dzikovska, G. Campbell, C. Callaway, N. Steinhauser, E. Farrow, J. Moore, L. Butler, and C. Matheson. 2008. Diagnosing natural language answers to support adaptive tutoring. In *International FLAIRS Conference*.

R. Freedman. 2000. Using a reactive planner as the basis for a dialogue agent. In *International FLAIRS Conference*.

Ken Hyland. 2007. Applying a gloss: Exemplifying and reformulating in academic discourse. *Applied Linguistics*, 28(2):266–285.

P. Jordan and M. A. Walker. 1996. Deciding to remind during collaborative problem solving: Empirical evidence for agent strategies. In *AAAI-96*, pages 16–23, August.

P. Jordan, B. Hall, M. Ringenberg, Y. Cui, and C.P. Rosé. 2007. Tools for authoring a dialogue agent that participates in learning studies. In *AIED 2007*.

P. Jordan, S. Katz, P. Albacete, M. Ford, and C. Wilson. 2012. Reformulating student contributions in tutorial dialogue. In *7th International Natural Language Generation Conference*, pages 95–99.

S. Kalyuga and P. Ayres. 2003. The expertise reversal effect. *Educational Psychology*, 38:23–31.

S. Katz and P. Albacete. 2013. A tutoring system that simulates the highly interactive nature of human tutoring. *Educational Psychology (Special Issue on Advanced Learning Technologies)*, 105(4):1126–1141.

D. Litman and K. Forbes-Riley. 2006. Correlations between dialogue acts and learning in spoken tutoring dialogues. *Natural Language Engineering*, 12(2):161–176.

D. Litman and S. Pan. 2000. Predicting and adapting to poor speech recognition in a spoken dialogue system. In *AAAI/IAAI*, pages 722–728.

S. A. McLeod. 2007. Stages of memory - encoding storage and retrieval. Retrieved from http://www.simplypsychology.org/memory.html.

D.S. McNamara, E. Kintsch, N.B. Songer, and W. Kintsch. 1996. Are good texts always better? text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, 14:1–43.

S. Murillo. 2008. The role of reformulation markers in academic lectures. In A.M. Hornero, M.J. Luzón, and S. Murillo, editors, *Corpus Linguistics: Applications for the Study of English*, pages 353–364. Peter Lang AG.

M.C. O'Connor and S. Michaels. 1993. Aligning academic task and participation status through revoicing: Analysis of a classroom discourse strategy. *Anthropology & Education Quarterly*, 24(4):318–335.

N. Person, A. Graesser, R. Kreuz, and V. Pomeroy. 2003. Simulating human tutor dialog moves in autotutor. *International Journal of Artificial Intelligence in Education*, 12(23-39).

M. A. Walker. 1996. The effect of resource limits and task complexity on collaborative planning in dialogue. *Artificial Intelligence Journal*, 85(1-2):181–243.