Out of the Lab and into the Classroom: An Evaluation of Reflective Dialogue in Andes

Sandra KATZ, John CONNELLY, and Christine WILSON
Learning Research and Development Center, University of Pittsburgh,
Pittsburgh, PA 15260

Abstract. Several laboratory studies have demonstrated the effectiveness of presenting students with "Reflection Questions" immediately after they have solved quantitative problems [1, 2]. The main goal of the two experiments described in this paper was to determine if the positive results of post-practice reflection from laboratory studies hold up in an actual classroom setting. We added prototype reflective dialogues to the Andes physics tutoring system [3, 4] and evaluated them in a course at the US Naval Academy. Consistent with prior research, the dialogues promoted conceptual understanding of physics. However, measures of retention and transfer to problem-solving ability showed only a marginal effect of completing the reflective dialogues, in the first experiment.

Keywords. Reflection, classroom-based research, natural-language interaction

Introduction

A well-known problem in the teaching of quantitative subjects such as physics is that even high-achieving students often leave a course with shallow knowledge [5, 6]. One technique that human tutors use to support students in tying a problem with its associated concepts, and in formulating abstract solution schemata from a set of similar problems, is to ask them qualitative questions immediately after they have solved a problem and to guide them in answering these questions. Such questions are commonly called "reflection questions" (RQs) [2], and the student-tutor discussions that stem from them are known as "post-practice reflective dialogues" (PPRs) [1].

Several studies which were conducted in a laboratory setting have shown that reflection questions and their ensuing PPRs support the development of conceptual knowledge and problem-solving ability [1, 2]. The main goal of the two "*in vivo*" experiments described in this paper was to determine if these results would hold up in an actual classroom setting. Toward this end, we developed prototype reflective activities for the Andes physics tutoring system [3, 4] and evaluated them in a physics course at the US Naval Academy.

Andes seemed like an ideal environment in which to couch our manipulation. Although Andes has been shown to significantly enhance students' problem-solving performance when compared with students in "traditional" sections of the course that do homework using paper-and-pencil, evaluations have not shown conclusively that the tutor enhances conceptual knowledge [4]. Furthermore, several physics instructors commented that they would like the tutor to do more to combat "shallow learning."

1. Overview of Andes

Andes is described in detail in [3] and summarized in [4]. We review Andes' essential features as background for our discussion of the reflective activities that we added to it.

Andes was designed to be used with most textbooks for first-year college physics courses, or for advanced high school courses. It provides over 500 problems for students to solve, just as they would with pencil and paper, but with the added benefit of coaching when they get stuck and immediate feedback on the correctness of their actions. Andes' coaching is designed to be increasingly directive, with general hints followed by more specific advice about what to do next and how to do it. Conceptual knowledge is mainly included in the top-level hints. However, many students gloss over these hints in order to get to the most detailed "bottom out" hints, which can lead to shallow learning [5]. We expected that students would be more receptive to instruction about physics concepts and problem-solving schemata *after* they solved problems than during problem solving, when they are intent on getting the correct answer. Andes is available for download at www.andes.pitt.edu.

2. Experiment 1

2.1 Background and Motivation

In addition to determining whether PPRs would support learning *in vivo*, the first experiment tested our prediction that the more *interactive* students are when responding to reflection questions, the more they will learn. This hypothesis is grounded in constructivist learning theory, as well as in empirical research that suggests that interaction—mutual contributions to dialogue by the student and the tutor—is a critical factor in the effectiveness of one-on-one human tutoring [7].

We compared three versions of reflective activity at different levels of interactivity. In each reflection condition, RQs were presented to students sequentially in a pop-up window as soon as they entered the correct solution to an Andes problem. This window was separate from the Andes interface, which remained visible in the background as students worked on the RQs.

Table 1 shows an RQ that was asked after a work-energy problem and an excerpt from the ensuing dialogue. "What if" questions such as this one were used throughout the intervention. In the least interactive form of reflection, the student entered a response to an RQ and then was shown an expert-generated response as feedback. The tutor then went on to the next question. We call this the canned-text response condition (CTR). It was designed to stimulate self-explanation, but not interaction.

The second reflection condition used Knowledge Construction Dialogues (KCDs) to guide students to a correct answer to each reflection question. KCDs are designed to guide students through Socratic-style questioning in learning one or a small number of concepts [8]. When a student enters an incorrect response to one of the tutor's questions, the dialogue engine invokes a remedial sub-dialogue. Table 1 shows a sample KCD (T = tutor; S = student). Lines 3-4 illustrate remediation. We refer to this form of reflection as the KCD condition.

Table 1. A sample reflective dialogue

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?

1. **T:** What quantities do you need to calculate work?
2. **S:** force, displacement
3. **T:** Unless you know both the force and displacement **vectors**, you also need the **angle** between the force and the displacement to calculate work.
4. **T:** 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. **S:** tension
6. **T:** Good. Does the tension force have a component in the direction of the suitcase's displacement?...

Although the KCD version of reflection is more interactive than the canned text version, it nonetheless does not support mixed initiative. For example, the student cannot ask a question. Our third reflection condition implemented a modest form of mixed initiative. At various points in the dialogue there were hyperlinks following the tutor's turn, which were associated with a question that a student might want to ask at that point in the dialogue. If the student selected one of these questions, the tutor answered the question and then returned to the main line of reasoning. We refer to this more interactive form of reflection as the mixed-initiative condition (MIX).

Based on the interactivity hypothesis, we predicted that students in the MIX condition would outperform students in the standard KCD condition, who would in turn outperform students in the CTR condition. In addition, we expected that students in any reflective condition would outperform students in a control condition (Ctrl) who used standard Andes without reflective dialogues.

2.2 Method

Our intervention consisted of 22 reflection questions, distributed across 9 problems in the work-energy (WE) unit of a first-year physics course at the US Naval Academy. The experiment was conducted in the fall of 2005 and lasted for about three weeks.

The experiment used a pre-test→intervention→post-test design. Both tests were administered to students in a laboratory setting. However, students did the intervention as homework, whenever they wished. The two tests were identical in form and content, with 15 multiple choice questions (12 qualitative and 3 quantitative WE problems).

Participants were 123 midshipmen (94 men, 22 women), enrolled in seven sections of General Physics I and taught by four instructors. They were randomly assigned to conditions within each course section. Conditions were balanced within category of academic major (Humanities, Engineering, Math/Science) and by sorted QPA within major category. There were 30-32 students per condition.

In order to measure retention and transfer, we analyzed copies of the hourly exam that followed the work-energy unit and of the final exam. The final exam was comprehensive; it contained seven work-energy problems. Both tests were strictly quantitative, thus allowing us to determine if there was a difference in the degree to which each condition supported problem-solving ability.

2.3 Results and Discussion

First, the bad news: Student participation in both homework problem solving and reflection was uneven across conditions and too low overall to allow us to reliably determine whether more interactive forms of reflection support learning better than less interactive forms. Among the 93 students assigned to one of the three treatment conditions, only 13 (14%) completed all 22 RQs prior to the post-test; 42 (45%) completed none of the RQs. The mean number of RQs completed was only 6.7.

Although it is possible that low participation of treatment subjects was due to dissatisfaction with the reflective dialogues, we believe that a more plausible explanation is that there was no course incentive (e.g., grade) for doing the dialogues. Many students did not do their homework, so they did not encounter the dialogues. Although instructors encouraged students to do the dialogues, they were not yet confident enough in the intervention to require students to do them.

Before doing a between-group comparison of student performance, we omitted data from students with a post-test duration of less than four minutes and a prevalence of “don’t know” selections. These students obviously were not trying to do well on the post-test. We also re-classified treatment subjects who did no dialogues as control subjects and combined the KCD and MIX conditions into one “dialogue” condition because very few MIX subjects asked follow-up questions. After these regroupings, we had 48 Ctrl, 17 CTR, and 38 dialogue subjects. Given the low participation in reflection across conditions, it is not surprising that we found no significant differences in post-test score or gain score—neither by ANCOVA nor by linear regression with major, Quality Point Average (QPA), and pre-test score as covariates.

Now for some good news: We were able to determine that having students engage in some form of reflective dialogue is better than no reflective activity. Using a yoked pairs analysis, we compared “treated” with “untreated” subjects—that is, students who responded to at least 5 RQs, in any reflection condition, with students who saw no RQs. We yoked subjects by pre-test scores and major, resulting in 19 pairs. As predicted, this analysis revealed that the mean gain score was significantly higher for treated than untreated subjects; M = 1.53, t(18) = 2.40, p = .03. Consistent with this result, a regression analysis treating number of RQs done as a continuous variable showed that this factor significantly predicted post-test score, as opposed to the number of problems that students completed; R^2 = .35, F(5,97) = 10.33, p < .0001.

The results of retention and transfer analyses were mixed. Using the same re-classification of subjects discussed previously, we found that students in the canned text condition did marginally better than students in the other three conditions on the seven work-energy problems on the final exam; F(2, 119) = 2.74, p = .07. For the hourly exams, we had data from only two course sections. Both exams contained only two work-energy problems. On one exam, regressing on the number of problems that students solved prior to the exam and the number of RQs that they completed showed a significant positive effect of the latter (p = .03), but only a marginal overall regression.

On the other exam, the same regression analysis showed a significant positive effect for the total number of problems done, but not for the number of RQs done. Again, the overall regression was marginal.

3. Experiment 2

Our second *in vivo* experiment took place a year later, in the fall of 2006. Our first goal was to verify the main result of the first experiment, which was that students who engaged in reflective activity after solving Andes problems outperformed control subjects who did not, as measured by post-test scores and gain scores over the pre-test. Secondly, we wanted to develop and test a reflective intervention that would enhance students' problem-solving ability, in addition to their conceptual knowledge.

A positive outcome of the first experiment was that it increased instructors' confidence in our intervention. Consequently, they requested a "Reflective Follow-up" module that covered most course units (not just work-energy), provided us with feedback on the reflective dialogues throughout the development process, and required treatment subjects to complete the dialogues. We used several methods to enforce participation, which we describe in the next section.

Our design of a new set of reflective dialogues that could potentially increase students' conceptual knowledge and problem-solving ability was motivated by research done by Dufresne, Gerace, Hardiman, & Mestre [9], who distinguish between three components of a problem-solving strategy: *what* a problem's main quantities, concepts and/or principles are; *how* to map principle applications to relevant equations, manipulate equations to solve for desired quantities, and extend manipulations into related contexts; and *why (not)* to apply a given principle or solve an alternate scenario via a different approach. Dufresne et al. [9] developed and tested a solution-planning tool (HAT, for *Hierarchical Planning Tool*) that guided students through each component. Although they found some improvements in problem-solving ability, they also observed that even high-achieving students were often unable to derive a correct set of equations using the tool. We speculated that HAT planning dialogues covered too much material, because they targeted all three knowledge components.

In light of these observations, we developed post-practice reflective dialogues that targeted mainly (but not solely) one component at a time and in order, within selected units of the course. The earliest assigned problem(s) in each unit contained a *what* reflective dialogue, middle problem(s) contained a *how* dialogue, and later problem(s) contained a *why* dialogue. We then capped off each unit with a pre-solution planning dialogue that integrated these components, as did HAT. We refer to these planning dialogues as "capstone dialogues."

This experiment compared students in two conditions. The treatment condition engaged in *what*|*how*|*why* reflective dialogues and capstone dialogues, implemented using standard, tutor-led KCDs. The control condition used standard Andes without KCDs, but solved more problems than did treatment subjects to control for time on task. We did not attempt to compare different versions of reflection in this experiment, as we did in the first experiment, partly because only three sections of the course participated (vs. seven sections during the previous fall), and we wanted to optimize the power of our analyses. Instead, we compared direct, explicit instruction in conceptual knowledge and solution planning with implicit instruction—that is, knowledge acquired through practice (via solution of extra homework problems).

3.2 Method

We again conducted this experiment *in vivo*—that is, within first-year physics classrooms at the US Naval Academy. Three sections of General Physics I, taught by two instructors, participated. Participants were 67 midshipmen (10 women, 57 men) who were randomly assigned to conditions within each course section. There were 33 treatment subjects and 34 control subjects. As before, conditions were balanced by academic major and by sorted QPA within each major category.

Our reflective intervention consisted of 21 post-practice dialogues, which targeted the *what*, *how*, and *why* components of problem solving in succession within selected course units. The intervention spanned five units (statics; translational dynamics, including circular motion; work-energy; power; and linear momentum, including impulse) lasting approximately eight weeks. In addition, there were five capstone planning dialogues (one toward the end of each unit). Control subjects solved five more problems than did treatment subjects.

As before, the experiment used a pre-test→intervention→post-test design and both tests were administered in a laboratory setting. Students solved the Andes problems as homework, on their own time. The two tests were identical in form and content, with 30 multiple choice questions (24 qualitative and 6 quantitative). To measure retention and transfer to problem-solving ability, we analyzed an hourly exam that covered several units, and the final exam. Both tests consisted solely of quantitative problems.

We implemented several strategies to ensure greater student participation in this experiment, relative to that of the first. With the course instructors' approval, Andes was modified so that students were required to do the target problems in a fixed sequence. (Ordinarily, students can do any problem in any order.) Students could not work on a problem until they solved all of its prerequisite problems and—if they were in the treatment condition—until they completed all of the dialogues associated with these problems. In addition, the KCDs were designed to discourage null responses. If students did not answer, the tutor prompted them to do so. We also tried to improve system recognition of student input by restating selected questions that a student answered incorrectly, giving the student another chance to answer via a multiple choice menu of correct and incorrect responses.

3.3 Results and Discussion

Analyses of pre- and post-test scores were more encouraging than for Experiment 1. After omitting scores from one student with a post-test duration of less than two minutes, we were left with treatment and control groups of equal size (n = 33). We also re-classified two treatment subjects who did no dialogues as control subjects (effective treatment and control *n*s = 31 and 35, respectively), although these subjects were not able to access the five extra control-group Andes problems. As we had hoped, ANOVA showed no significant differences between the effective treatment and control groups on pre-test score (respective *M*s = 12.10 and 11.77; $F < 1$), but treatment subjects had higher mean post-test scores (17.97 vs. 15.57; $F(1, 64) = 4.89$, $p = .031$), mean raw gain scores (5.87 vs. 3.80; $F = 5.62$, $p = .021$), and mean Estes gain scores (0.330 vs. 0.208; $F = 6.74$, $p = .012$) than did control subjects. In short, without regard to problem-solving practice, subjects who did KCDs did significantly better on the post-test than those who did not.

Student participation in both homework problem solving and dialogues was much improved over Experiment 1, with 21 of 34 control subjects (62%) and 23 of 33 treatment subjects (70%) finishing at least 80% of the assigned target problems (of 31 for control, 26 for treatment) prior to the post-test and with 25 treatment subjects (76%) finishing at least 80% of the 26 associated KCDs. However, participation was still far from perfect; 7 treatment subjects (21%) completed half or fewer of the assigned KCDs on time, and 2 of them completed no target problems or KCDs on time. We again treated KCD completion and target problem completion as continuous independent variables in regression analyses. Regressing post-test score on pre-test score, QPA, number of KCDs completed, and number of target problems completed ($R^2 = .52$, $F(4, 61) = 16.70$, $p < .00001$) showed positive contributions of all factors, but only pre-test score, QPA, and KCD completion were statistically significant (*p*s < .001, .05, & .05, respectively); problem completion was *ns* ($p = .54$). Therefore, across all subjects it was KCD completion, as opposed to target homework problem completion, that significantly predicted post-test performance.

To measure retention and transfer, we also analyzed scores from an hourly exam administered by one instructor to his two sections (*n*=47). All problems on this exam were quantitative and covered a subset of the units targeted during the intervention. Scores on this exam ranged from 176 to 382 (of 400), and were highly correlated with pre- and post-test scores; r*s*(45) = .54 and .71, *p*s < .0001 and .00001. However, ANOVAs showed no differences between groups (*F*s < 1), and regression of subscores on QPA, KCDs completed, and target problems completed ($R^2 = .54$, $F(3, 43) = 16.58$, $p < .00001$) showed positive contributions of all factors but a significant effect of only QPA ($p < .00001$). Therefore, student performance on the hourly exam was not significantly affected by either KCD completion or target problem completion.

In summary, student performance on the post-test relative to the pre-test was significantly influenced by the number of dialogues they completed, as opposed to the number of target problems they completed prior to the post-test. However, neither measure had a significant effect on student performance on an exam that focused on problem-solving ability. Analyses of the final exam data are in progress.

4. Conclusion

The two studies described in this paper suggest that the positive results of reflection from laboratory studies hold up in an actual classroom setting. Although our first experiment did not allow us to test our hypothesis that more interactive forms of reflection are more beneficial than less interactive forms, such as canned text, we found some evidence against this hypothesis. For example, in the first experiment, the canned text condition marginally outperformed the other three conditions on the final exam. Several studies comparing human or automated dialogue with canned text have also failed to show an advantage of the former [1, 8].

Analyses of retention and transfer to problem-solving ability did not reveal a significant effect of the reflective dialogues in either experiment. We had hoped that the capstone dialogues that treatment students completed before selected problems, at the end of each unit, would support problem-solving ability. It is possible that the intervention consisted of too few capstone dialogues (only five) to observe any effect.

What seems to be the critical feature of post-practice reflection in supporting conceptual understanding is the explicit instruction that it provides in domain knowledge. PPR may help students to fill in knowledge gaps, resolve misconceptions, and abstract from the case at hand so that they are better prepared to engage in constructive activity (e.g., self-explanation) in future problems. Preliminary research comparing Andes (which encourages implicit learning of problem-solving strategies) with Pyrenees, a system that teaches problem-solving strategies explicitly, also suggests an advantage for explicit instruction [10]. Further research is needed to identify the mechanisms that drive learning from reflective dialogue, and to increase its potential to enhance problem-solving ability in addition to conceptual knowledge.

Acknowledgements. This research was supported by grants from the Office of Naval Research (grant number N00014-97-1-0848) and the Pittsburgh Science of Learning Center, which is funded by the National Science Foundation award number SBE-0354420. The data and views expressed are not necessarily endorsed by these agencies. We appreciate the collaboration of Professors Robert Shelby, Donald Treacy, and Mary Wintersgill throughout the development and testing of reflective dialogues. Kurt VanLehn offered many valuable suggestions. Pamela Jordan and Anders Weinstein provided crucial software support.

References

[1] 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*, *13* (1), 79-116.

[2] Lee, A. Y., & Hutchison, L. (1998). Improving learning from examples through reflection. *Journal of Experimental Psychology: Applied, 4* (3), 187-210.

[3] 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. *International Journal of Artificial Intelligence and Education*, *15* (3).

[4] 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: Five years of evaluations. In G. McCalla, C. K. Looi, B. Bredeweg & J. Breuker (Eds.), *Artificial Intelligence in Education* (pp. 678-685). Amsterdam, Netherlands: IOS Press.

[5] Aleven, V., & Koedinger, K. R. (2002). An Effective Meta-cognitive Strategy: Learning by Doing and Explaining with a Computer-Based Cognitive Tutor. *Cognitive Science, 26* (2), 147-179.

[6] Halloun, I.A., & Hestenes, D. (1985). The initial knowledge state of college students. *The American Journal of Physics, 53*, 1043-1055.

[7] Chi, M. T. H., Siler, S., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science*, *25*, 471-533.

[8] Rosé, C. P., Bhembe, D., Siler, S., Srivastava, R. & VanLehn, K. (2003). Exploring the effectiveness of knowledge construction dialogues. In U. Hoppe, F. Verdejo, & J. Kay (Eds.), *Proceedings of the 11th International Conference on Artificial Intelligence in Education, AI-ED 2003* (pp. 497-499). Amsterdam: IOS Press.

[9] Dufresne, R.J., Gerace, P., Hardiman, T., & Mestre. (1992). Constraining novices to perform expertlike analyses: Effects on schema acquisition. The Journal of the Learning Sciences, 2 (3), 307-31.

[10] VanLehn, K., Bhembe, D., Chi, M., Lynch, C., Schulze, K., Shelby, R., Taylor, L., Treacy, D., Weinstein, A., & Wintersgill, M. (2004). Implicit versus explicit learning of strategies in a non-procedural cognitive skill. In J. C. Lester, R. M. Vicari, & F. Paraguacu, (Eds.), *Intelligent Tutoring Systems: 7th International Conference* (pp. 521-530). Berlin: Springer-Verlag Berlin & Heidelberg GmbH & Co.