When is Tutorial Dialogue More Effective Than Step-based Tutoring?

Min Chi[1], Pamela Jordan[2], and Kurt VanLehn[3]

[1] Computer Science Department, North Carolina State University,Raleigh NC USA,mchi@ncsu.edu,
[2] Learning Research and Development Center, University of Pittsburgh, Pittsburgh, PA USA pjordan@pitt.edu
[3] School of Computing, Informatics and Decision Science Engineering, Arizona State University, AZ USA, Kurt.Vanlehn@asu.edu

Abstract. It is often assumed that one-on-one dialogue with a tutor, which involves micro-steps, is more effective than conventional step-based tutoring. Although earlier research often has not supported this hypothesis, it may be because tutors often are not good at making micro-step decisions. In this paper, we compare a micro-step based NL-tutoring system that employs induced pedagogical policies, Cordillera, to a well-evaluated step-based ITS, Andes. Our overall conclusion is that the pairing of effective policies with a micro-step based system does significantly outperform a step-based system; however, there is no significant difference in the absence of effective policies. Moreover, while micro-step tutoring is more time-consuming, the findings still hold for five out of six learning performance measures when time on task is factored out.

Keywords: Natural Language tutoring systems, Step-based tutoring systems, Reinforcement Learning, Pedagogical Strategy

1 Introduction

In domains like math and science solving a problem requires producing an argument, proof or derivation consisting of one or more inference steps, and each step is the result of applying a domain principle, operator or rule. For instance, solving an algebraic equation, 2x+5=21 can be done via two steps: 1) subtract the same term 5 from both sides of the equation; and 2) divide both sides by 2.

A typical Intelligent Tutoring System (ITS) is step-based [13] while human tutors or Natural Language (NL) tutoring systems, by contrast, often scaffold students via a series of micro-steps leading to the full step. A step corresponds to (writing) actions that are ordinarily done even without the tutor, whereas micro-steps correspond roughly to dialogue acts in tutorial dialogues. For example, in order to solve a physics problem, the student needs to apply several domain principles, some of which may need to be applied multiple times. Each principle application can be seen as a step in the ITS. In a physics tutor, for example, applying the definition of Kinetic Energy ($KE = \frac{1}{2}mv^2$) to solve for the kinetic

1. **T:** So let's start with determining the value of KE_0.
2. **T:** Which principle will help you calculate the rock's kinetic energy at T0? Please provide the name of the principle, not an equation. **{ELICIT}**
3. **S:** Definition of kinetic energy
4. **T:** Yes, I agree. Now I will write the equation for applying the definition of kinetic energy to the rock at T0: KE0 = 1/2*m*v0^2 **{TELL}**

Fig. 1. Illustrations of Two Micro-Steps

energy of a falling rock at T_0 is a step; once a student enters a step, then the tutor gives feedback and/or hints. Human tutors, by contrast, often scaffold students via a series of micro-steps leading to the full step. In the step mentioned above, for instance, a human tutor can take the following **micro-level** steps: selecting the principle to apply; writing the corresponding equation; solving the equation; and engaging in some qualitative discussion about the principle.

Fig. 1 illustrates two micro-steps and each numbered line represents a dialogue turn. The labels **T** and **S** designate tutor and student turns respectively. In this example, the tutor and the student first select a principle (lines 2 & 3) and then write the corresponding equation (line 4). Some of the tutor turns in Fig. 1 are labeled **{ELICIT}** or **{TELL}**. This label designates a ***tutorial decision step*** wherein the tutor has to make a tutorial decision whether to ask the student for the requisite information or to tell it to the student. For example, in line 2, the tutor chooses to *elicit* the answer by asking, "Which principle will help you calculate the rock's kinetic energy at T0? Please provide the name of the principle, not an equation." If the tutor elects to tell, however, then he or she would state, "To calculate the rock's kinetic energy at T0, let's apply the definition of Kinetic Energy."

One common hypothesis as to the effectiveness of human one-on-one tutoring comes from the detailed management of "micro-steps" in tutorial dialogue[6, 7] and thus suggests that micro-step based tutors are more effective than step-based tutors. In several tests of this hypothesis, however, neither human tutors nor NL tutors designed to mimic human tutors, outperformed step-based tutors once content was controlled to be the same across all conditions [5, 12]. All three types of tutors were more effective than no instruction (e.g., students reading material and/or solving problems without feedback or hints). One possible conclusion is that tutoring is effective, but the micro-steps of human tutors and NL tutoring systems provide no additional value beyond conventional step-based tutors[13].

Alternatively, we argue that the lack of difference between micro-step and step-based tutors is because neither the human tutors nor the NL tutoring systems involved in those studies were good at making micro-step decisions and several studies provide some support for this claim[3, 11, 2]. Previously, we investigated the impact of pedagogical policies on student learning by comparing different versions of a micro-step based NL tutoring system called Cordillera [2].

We applied a general data-driven methodology, Reinforcement Learning (RL), to induce pedagogical policies directly from student interactivity logs and found that Cordillera with effective pedagogical policies, RL-induced Cordillera significantly out-performed other versions of Cordillera. However, it is still unclear whether the former is significantly better than a step-based ITS.

In this paper, we directly compare RL-induced Cordillera with a well-evaluated step-based conventional ITS, Andes [14]. Our main research question is: *Can a NL tutoring system with machine-learned pedagogical policies be more effective than a step-based ITS?* Overall, we find that RL-induced Cordillera significantly outperforms Andes. In order to investigate whether this result is indeed caused by effective RL-induced policies, we also compare Andes to two other versions of Cordillera: Hybrid-RL and Random. In the following, we will briefly describe the two types of tutoring systems and the pedagogical policies employed in them and then describe our study and finally present our results.

2 Two Types of ITSs

The Micro-step based Cordillera: NL Tutorial Dialogue System
The Cordillera tutorial dialogue system tutors students in both quantitative and qualitative physics in the work-energy domain and was implemented using the TuTalk tutorial dialogue system toolkit toolkit [8]. TuTalk supports dialogues in which a tutor tries to elicit the main line of reasoning from a student by a series of coherent questions. This style of dialogue was inspired by CIRCSIM-Tutor's directed lines of reasoning [5]. The Cordillera style of dialogue is system-initiative in that the system always chooses the topics discussed.

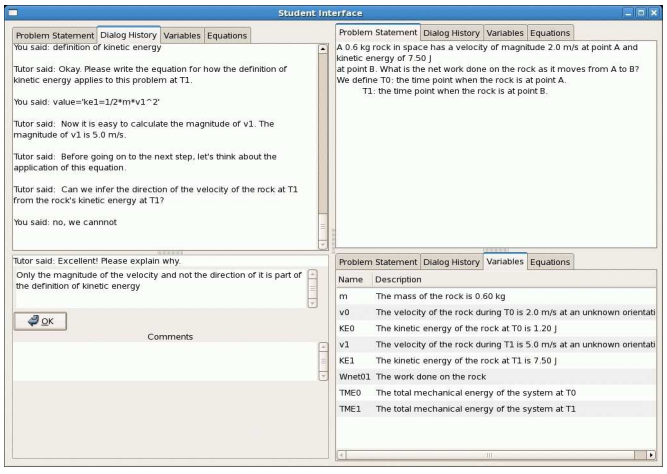

Fig. 2. An example of the Cordillera interface

Figure 2 illustrates a sample student dialogue with Cordillera. The upper top right pane of the figure shows the problem that the student is attempting to solve. The top left pane shows a portion of the dialogue history, and illustrates a few questions and student responses, as well as a number of system informs; the pending tutor question is shown in the input pane at the bottom followed by the response the student is entering. Finally, the variables in the bottom right pane and the equations (hidden) were entered either by the student using a form interface (not shown) or provided by the tutor. When the tutor asks the student to compute the value for a variable, the student must transform the equation to a solvable form with the known values substituted and then the tutor will do the final calculation. In order to avoid confounds due to imperfect NL understanding, a human wizard replaced the NL understanding module. During tutoring, the wizard matched students' answers to one of the available responses but made no tutorial decisions.

The step-based Andes Tutoring System
Andes provides a multi-paned screen that consists of a problem-statement window, a variable window, an equation window, and a dialogue window. An example of the Andes interface, as the student would see it, is shown in Figure 3. On Andes, students construct and manipulate a solution. The interaction is open-ended, event-driven and student-initiated. Students can enter an equation that is the algebraic combination of several principle applications and Andes provides immediate feedback on each entry. Andes can also algebraically manipulate equations to calculate the value for a variable. It considers an entry correct if it is true, regardless of whether it is useful for solving the problem. When an entry is incorrect, students can either fix it independently, or ask for what's-wrong help. When they do not know what to do next, they can ask for next-step help. Both next-step and what's-wrong helps are provided via a sequence of hints that gradually increase in specificity. The last hint in the sequence, called the bottom-out hint, tells the student exactly what to do.

Andes provides conceptual and procedural help that is designed to encourage students to think on their own. Students can always enter any correct step and Andes does not attempt to determine their problem-solving plans. If necessary for giving a hint, it asks students what principle they are working on. If students indicate a principle that is part of a solution to the problem, Andes hints an uncompleted step from the principle application. If no acceptable principle is chosen, Andes picks an unapplied principle from the solution that they are most likely to be working on.

3 Decision Policies within Cordillera and Andes

In many tutoring systems, the system's behaviors can be viewed as a sequential decision process wherein, at each discrete step, the system is responsible for selecting the next action to take. Pedagogical strategies are defined as policies to decide the next system action when multiple are available. Each of these sys-

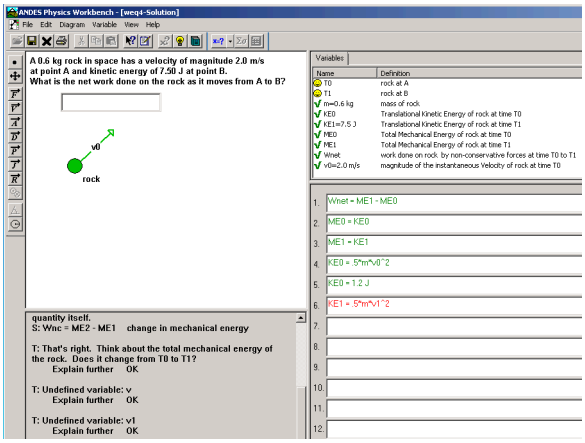

Fig. 3. An example of the Andes interface

tem decisions affects the user's successive actions and performance. Its impact on student learning cannot often be observed immediately and the effectiveness of one decision also depends on the effectiveness of subsequent decisions. Ideally, an effective tutor should craft and adapt its decisions to users' needs [1, 10]. However, there is no existing well-established theory on how to make these system decisions effectively. In this work, different versions of micro-step based Cordillera employed different pedagogical policies. The step-based Andes employs hand-coded rules.

Three versions of Cordillera - Random, Hybrid-RL, and RL-induced - were involved. The only difference among the three is the policy used. Random Cordillera made tutorial decisions randomly. Hybrid-RL Cordillera used expert-guided data-driven induced rules. These rules were induced by using 18 features and a greedy-like procedure to prune the features to meet efficiency and training constraints[4]. Both the initial features and pruning procedure were suggested by human experts and the final induced policies were also checked and approved by human experts. But no significant difference was found on overall learning performance between the Hybrid-RL and random policies. For RL-induced Cordillera, the data-driven approach was greatly improved. More specifically, the RL approach involved a much larger feature set (50 features), and more advanced domain-general feature selection approaches. Human experts were not involved in directing the policy generation. As reported earlier[2], these RL-induced policies indeed helped students learn more and in a deeper way than either Hybrid-RL or random policies.

Andes, on the other hand, like most existing ITSs employs hand-coded pedagogical policies. For example, help in Andes is provided upon request because it is assumed that students know when they need help and will only process help when they desire it. A student deciding to request help can be seen as a human-like decision policy for whether to skip or not skip content.

4 Methods

Participants: A total of 163 participants used either Andes or one of the three versions of Cordillera: the Andes group comprised 33 students; the Random Cordillera Group comprised 64 students so that we could collect enough data for RL policy induction; the Hybrid-RL Cordillera Group comprised 37 students; and the RL-induced Cordillera group comprised 29 students. All participants were recruited in the same way but in different years.

Domain & Procedure: The training covered the first-year college physics work-energy domain. All participants experienced identical procedures: 1) a background survey; 2) read a textbook covering the target domain knowledge; 3) took a pretest; 4) solved the same seven training problems in the same order on either Andes or Cordillera; and 5) finally took a posttest. The pretest and posttest were identical and contained 16 quantitative items and 16 qualitative items. Both quantitative and qualitative items include multiple choice and open-ended problems.

Students' learning outcomes were measured by using three types of scores: quantitative, qualitative and overall. All tests were graded in a double-blind manner by experienced graders. In a double-blind manner, neither the students nor the graders know who belongs to which group. For comparison purpose all test scores were normalized to fall in the range of [0,1].

Except for following the policies (Random, Hybrid-RL, or RL-induced), the remaining components of Cordillera, including the interface, the training problems, and the tutorial scripts, were identical for all students. However, there are some noticeable differences for the Andes training compared to Cordillera.

Differences in the Training: The Cordillera dialogues guided students through the training problems by hinting at the next problem solving step to be completed, or telling them what it is. Hints took the form of short answer questions. In addition to guiding the student through problem solving, Cordillera also attempted to help the student increase his/her conceptual understanding of the domain by asking for justifications for the most important problem solving steps. The decision for when to ask for a justification was determined by a set of pedagogical policies. For an example of a justification requested during problem solving, see the current tutor turn in the bottom left input pane in Figure 2. There was also a post-problem discussion for each problem which sought to increase the student's conceptual qualitative understanding.

We implemented the same seven training problems in Andes and because Cordillera provided drawings and pre-defined some variables for each problem, we set-up Andes to provide the same. We added a post-problem discussion to Andes by collecting all the post-problem discussion for Cordillera into a *static* text document so that the content coverage for post-problem discussion was about the same. The post-problem discussion was delivered in a series of web pages after the experimenter verified that the student had completed the Andes problem.

Note that we did not attempt to provide identical content for the problem solving help since it reflects two different tutoring systems, but what is available is similar. For example, while the Cordillera system's micro-steps will always present the content illustrated in Fig. 1, Andes will show the following series of hints for this same step after the student makes four consecutive help requests: 1) Why don't you continue with the solution by working on the definition of kinetic energy. 2) What is the kinetic energy of the rock at T0? 3) The kinetic energy of an object is defined as one half its mass times its velocity squared. That is, $0.5 * m * v^2$. 4) Write the equation KE0 $= 0.5 * m * v0^2$. So for this illustration asking for all hints on the Andes step is equivalent to a decision to tell for all the related micro-steps in Cordillera.

While the problem solving help content is similar, there is also some conceptual qualitative discussion during Cordillera's problem solving that Andes does not offer. It is up to the student to consider the concepts involved on their own. However, as has been pointed out, novice students have a tendency to simply manipulate equations to isolate the unknown and seldom consider the conceptual knowledge involved during problem solving [9].

5 Results

Overall Training Time

A one-way ANOVA showed significant differences among the four groups on overall training time: $F(3, 154) = 53.90$, $p < 0.001$. The Andes group spent significantly less time[4] than the other three groups but there were no significant differences in time on task among the three Cordillera groups. The average training time (in minutes) across the seven training problems, was $M = 115.94$, $SD = 42.03$ for Andes, $M = 280.38$, $SD = 66.88$ for Random, $M = 294.33$, $SD = 87.51$ for Hybrid-RL, and $M = 259.99$, $SD = 59.22$ for RL-induced.

Learning Performance

Although students were recruited during different time periods, they appear balanced on incoming competence across the conditions. A one-way ANOVA showed that there were no significant differences in pretest scores among the four groups on either quantitative: $F(3, 159) = 1.18$, $p = .32$, or qualitative: $F(3, 159) = 0.06$, $p = .98$, or overall questions $F(3, 159) = 0.46$, $p = .71$.

A repeated measures analysis using test (pretest vs. posttest) as a factor and test score as the dependent measure showed that there was a main effect for test. All four groups of students scored significantly higher on the posttest than the pretest, $F(1, 32) = 19.87$, $p < 0.001$ for Andes, $F(1, 63) = 78.37$, $p < 0.001$ for Random, $F(1, 36) = 48.36$, $p < 0.001$ for Hybrid-RL, and $F(1, 28) = 238.58$, $p < 0.001$ for RL-induced.

The same results were found from pretest to posttest on both quantitative and qualitative questions as well. More specifically, on quantitative questions,

[4] Some reading times for the last problem were lost so we used the minimum average reading time for all other easier problems.

Table 1. RL-induced Cordillera vs. Andes on Various Test Scores

Test Item Set	Test Score	RL-induced Cordillera	Andes	Stat	cohen d
quant	Pre	0.35 (0.25)	0.28 (0.26)	$t(60) = 1.01, p = .28$	0.27
	Post	0.64 (0.22)	0.41 (0.30)	$t(60) = 3.29, p = 0.002$	$0.87 **$
	Adj Post	0.61 (.18)	0.44 (.17)	$F(1, 59) = 13.793, p < .0001$	$0.97 **$
	NLG	0.49 (0.28)	0.16 (0.38)	$F(1, 59) = 14.442, p < 0.0001$	$0.99 **$
qual	Pre	0.46(0.12)	0.45(0.14)	$t(60) = 0.40, p = .688$	0.08
	Post	0.65 (0.14)	0.54 (0.18)	$t(60) = 2.68, p = 0.010$	$0.68 **$
	Adj Post	0.65 (.14)	0.54 (.14)	$F(1, 59) = 7.74, p = .007$	$0.79 **$
	NLG	0.36 (0.24)	0.14 (0.34)	$F(1, 59) = 8.86, p = 0.004$	$0.75 **$
Overall	Pre	0.42 (0.15)	0.39 (0.16)	$t(60) = 0.87, p = .39$	0.19
	Post	0.65 (0.15)	0.50 (0.21)	$t(60) = 3.35, p = 0.001$	$0.82 **$
	Adj Post	0.64 (.11)	0.51 (.12)	$F(1, 59) = 16.50, p < .0001$	$1.13 **$
	NLG	0.42 (0.19)	0.17 (0.28)	$F(1, 59) = 15.97, p < 0.0001$	$1.04 **$

$F(1, 32) = 15.83$, $p < 0.001$ for Andes, $F(1, 63) = 33.55$, $p < 0.001$ for Random, $F(1, 36) = 58.01$, $p < 0.001$ for Hybrid-RL, and $F(1, 28) = 95.79$, $p < 0.001$ for RL-induced. On qualitative questions, $F(1, 32) = 7.68$, $p = 0.009$ for Andes, $F(1, 63) = 40.62$, $p < 0.001$ for Random, $F(1, 36) = 17.20$, $p < 0.001$ for Hybrid-RL, and $F(1, 28) = 89.56$, $p < 0.001$ for RL-induced. Therefore all four conditions made significant gains from pre-test to post-test across all three sets of questions: quantitative, qualitative and overall questions. In order to investigate whether micro-step based tutors can be more effective than step-based tutors, we first investigated whether the most effective version of Cordillera would outperform Andes.

RL-induced Cordillera vs. Andes

Table 1 compares the pre-test, post-test, adjusted post-test, and NLG scores between the RL-induced Cordillera and Andes conditions by question type. The adjusted Post-test scores were compared between the two conditions via an ANCOVA with the corresponding pre-test score as a covariate. NLG measures students' gain *irrespective of their incoming competence*: $NLG = \frac{posttest - pretest}{1 - pretest}$. Here 1 is the maximum score. The third and fourth columns in Table 1 list the means and SDs of the two groups' corresponding scores. The fifth column lists the statistical comparison and the last column lists the effect size of the comparison using Cohen's d[5]. Table 1 shows that there was no significant difference between the two conditions on pre-test scores. However, there were significant differences between them on the post-test, adjusted post-test, and NLG scores for all three question types.

We then compared the two groups' performance on six types of learning measures: {Quantitative, Qualitative, Overall} × {Posttest, NLG} using both

[5] Which is defined as the mean learning gain of the experimental group minus the mean of the control group, divided by the groups' pooled standard deviation.

pre-test and total training time as the covariates. On one measure, quantitative posttest, there was no significant difference between the two groups: $F(1, 58) = 2.34, p = 0.132$. But on the remaining five measures, RL-induced Cordillera significantly outperformed Andes: $F(1, 58) = 7.27, p = 0.009$ for qualitative posttest, $F(1, 58) = 5.94, p = 0.018$ for overall posttest, $F(1, 59) = 4.72, p = 0.034$ for quantitative NLG, $F(1, 59) = 7.34, p = 0.009$ for qualitative NLG and $F(1, 58) = 9.71, p = 0.003$ for overall NLG respectively.

In sum, our results showed that micro-step based tutors can indeed be more effective than step-based tutors as RL-induced Cordillera significantly outperformed Andes on all types of test questions. Even when time on task is factored out, the same results hold for five out of six learning measures. Next, we compared Random and Hybrid-RL Cordillera with Andes to investigate whether the micro-step tutor would still be more effective than the step-based tutor *without* effective pedagogical policies.

Random vs. Andes & Hybrid-RL Cordillera vs. Andes: There were no significant differences between the Random-Cordillera and Andes groups on any of the learning outcome measures. Since Andes students spent significantly less time than Cordillera students, we compared the two conditions' posttest scores using both pre-test score and total training time as covariates and their NLG scores using total training time as the covariate. To our surprise, we still found no significant differences between the two groups. We had expected the efficiency of the Andes group to have some impact.

Similar results were found when we compared Hybrid-RL Cordillera and Andes on all types of learning outcome measures either when time on task is factored in or out. Since Hybrid-RL Cordillera employed human-influenced pedagogical rules, these results again indicate that expert tutors' pedagogical rules may not always be effective. Again, this study suggests that fine-grained interactions at micro-steps are a potential source of pedagogical power, but human tutors may not be particularly skilled at choosing the right micro-steps.

6 Conclusions and Future Work

Although it is often believed that micro-step based NL tutoring systems should be more effective than conventional step-based ITSs, little evidence was previously found to support this. Our hypothesis is that it is because the existing micro-step based NL tutoring systems do not employ effective pedagogical strategies. Previous work applied a general data-driven RL approach to induce effective pedagogical policies directly from student logs and found them to be more effective than either random or Hybrid-RL policies. However, it was still not clear whether these RL-induced policies would make micro-step based NL tutoring systems more effective than step-based ITSs.

In this paper, we found that RL-induced Cordillera significantly outperforms Andes while neither Hybrid-RL Cordillera nor Random Cordillera were significantly different from step-based Andes. Our overall conclusion is that a micro-step based system with effective RL-induced policies can significantly outperform

a step-based ITS with hand-coded policies; however, there is no significant difference between micro-step based and step-based tutoring systems in the absence of effective policies. Note that micro-step based Cordillera is more time-consuming than Andes. However, even when time on task is factored out, the micro-step based tutoring system with effective RL-induced policies is still significantly better than the step-based tutoring systems with hand-coded policies on five out of six learning performance measures.

Future work that remains is to explore policy-induction for Andes and to conduct a comparison of step-based tutoring to micro-step tutoring when both have effective RL-induced pedagogical policies. This may improve our understanding of the grain-size (step vs. micro-step) issue.

Acknowledgments This work was supported by NSF Award #0325054.

References

1. Anderson, J.R., Corbett, A.T., Koedinger, K.R., Pelletier, R.: Cognitive tutors: Lessons learned. The Journal of the Learning Sciences 4(2), 167–207 (1995)
2. Chi, M., VanLehn, K., Litman, D.J., Jordan, P.W.: Empirically evaluating the application of reinforcement learning to the induction of effective and adaptive pedagogical strategies. User Model. User-Adapt. Interact. 21(1-2), 137–180 (2011)
3. Chi, M.T.H., Siler, S., Jeong, H.: Can tutors monitor students' understanding accurately? Cognition and Instruction 22(3), 363–387 (2004)
4. Chi, M., Jordan, P.W., VanLehn, K., Litman, D.J.: To elicit or to tell: Does it matter? In: Dimitrova, V., Mizoguchi, R., du Boulay, B., Graesser, A.C. (eds.) AIED. pp. 197–204. IOS Press (2009)
5. Evens, M., Michael, J.: One-on-one Tutoring By Humans and Machines. Mahwah, NJ: Erlbaum (2006)
6. Graesser, A.C., Person, N., Magliano, J.: Collaborative dialog patterns in naturalistic one-on-one tutoring. Applied Cognitive Psychology 9, 359–387 (1995)
7. Graesser, A.C., VanLehn, K., Rosé, C.P., Jordan, P.W., Harter, D.: Intelligent tutoring systems with conversational dialogue. AI Magazine 22(4), 39–52 (2001)
8. Jordan, P.W., Hall, B., Ringenberg, M., Cui, Y., Rosé, C.: Tools for authoring a dialogue agent that participates in learning studies. In: Proceedings of AIED 2007. pp. 43–50 (2007)
9. Leonard, W., Dufresne, R., Mestre, J.: Using qualitative problem-solving strategies to highlight the role of conceptual knowledge in solving problems. American Journal of Physics 64(12) (1996)
10. Phobun, P., Vicheanpanya, J.: Adaptive intelligent tutoring systems for e-learning systems. Procedia - Social and Behavioral Sciences 2(2), 4064 – 4069 (2010), innovation and Creativity in Education
11. Putnam, R.T.: Structuring and adjusting content for students: A study of live and simulated tutoring of addition. Amer. Edu. Res. Journal 24(1), 13–48 (1987)
12. VanLehn, K., Graesser, Jackson, Jordan, Olney, Rose: When are tutorial dialogues more effective than reading? Cog. Sci. 31(1), 3–62 (2007)
13. VanLehn, K.: The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist 46(4), 197–221 (2011)
14. VanLehn, K., Lynch, C., Schulze, K., Shapiro, J.A., Shelby, R., Taylor, L., Treacy, D., Weinstein, A., Wintersgill, M.: The andes physics tutoring system: Lessons learned. IJAIED 15(3), 147–204 (2005)