Scaffolding Problem Solving with Annotated, Worked-Out Examples to Promote Deep Learning

Michael A. Ringenberg and Kurt VanLehn

University of Pittsburgh, Learning Research and Development Center
3939 O'Hara St., Pittsburgh PA 15260, USA
412-624-3353
mringenb@pitt.edu, vanlehn@cs.pitt.edu
http://www.pitt.edu/~mringenb/

Abstract. This study compares the relative utility of an intelligent tutoring system that uses procedure-based hints to a version that uses worked-out examples for learning college level physics. In order to test which strategy produced better gains in competence, two versions of Andes were used: one offered participants graded hints and the other offered annotated, worked-out examples in response to their help requests. We found that providing examples was at least as effective as the hint sequences and was more efficient in terms of the number of problems it took to obtain the same level of mastery.

1 Introduction

At the heart of most educational research is the search for ways to improve the instruction of novices. One strategy that has been found to be very effective is one-on-one human tutoring [1]. The economics of providing one-on-one tutoring has prompted the investigation of other techniques to boost learning. Another technique is to use intelligent tutoring systems to supplement classroom instruction and to substitute for individualized instruction. Another technique is to use embedded examples in instructional material [2], [3], [4], [5], [6]. As both paths have met with some success, it is worth comparing them and exploring ways to combine them.

Our study was done with modification of Andes, an intelligent tutoring system that aids the instruction of college-level introductory physics [7]. The main function of Andes is to present students with problems and to let the students solve them with the option of receiving adaptive scaffolding from the system. The two types of adaptive scaffolding in Andes are flag feedback and hints. Flag feedback marks the student's input as either correct or incorrect. When the student asks for help, Andes presents the student with a hint. The hint either points out what is wrong with the input or suggests a step to do next. The hint is based on the anticipated next step in solving the problem. It is designed to help the student identify and apply the missing relevant basic principles and definitions.

In this way, Andes tries to link the current problem-solving step with facts the student has already been taught.

Each hint is staged in a graded fashion known as a hint sequence. The student is typically presented first with a vague suggestion to prompt self explanation of the next step or to identify and correct the current error. The student can then ask for the next level in the hint sequence if the student judges that the previous hint was insufficient. The hints become more concrete as the sequence is followed. The last level in a hint sequence typically supplies the entire anticipated next correct problem-solving step. This is referred to as the bottom-out hint. This graded structure of hints has been used in several intelligent tutoring systems (For more information on Andes, please see `http://www.andes.pitt.edu/`). Students can and do resort to "help abuse" when this form of adaptive scaffolding is offered [8]. Students can click through the hints rapidly in order to get to the bottom-out hint and will ignore the rest of the hint sequence. This strategy is a problem because it is associated with shallow learning [8].

Our basic hypothesis is that novice students will learn more effectively if we replace Andes' hint sequences with worked-out near-transfer examples. A worked-out example is a solved problem with all of the anticipated problem-solving steps explicitly stated. A near-transfer example has a deep structure similar to that of the current problem and uses the same basic principles. Several lines of evidence suggest that worked-out examples will be more effective for novices than hint sequences.

First, based on an observation from previous Andes studies, some students will find the solution to one problem through help abuse and then refer back to that solved problem when faced with a similar problem. In essence, they are using the first problem to create a worked-out example. This observation is consistent with studies showing that novices prefer to learn from examples as opposed to procedural instructions [9].

Second, we suspect that the hints provided by Andes can provide good targeted help to students who are already familiar with the subject material and have an adequate understanding of the underlying principles. However, for novices, the first hints in the graded hint sequence probably make little sense. Novices are not sufficiently familiar with the subject material for the hints to activate the reasoning needed to finish the anticipated next step, nor are they familiar enough with the problem solving structure to understand why the hint is relevant.

Third, worked-out examples have been shown to be effective instruction in some cases. In one study, worked-out examples were more effective than presenting procedural rules [3]. However, examples are more effective when they alternate with problem solving, presumably because studying large blocks of examples becomes boring [10]. By using a single example in place of a hint sequence for each problem, we can avoid the boredom of large example blocks.

On the other hand, worked-out examples are not always effective. Their usefulness requires that students self-explain the solution steps listed in the example. A self-explanation for an example is a meaningful and correct explanation of a step in the student's own words [11]. Unfortunately, students do not tend to produce self-explanations spontaneously and many students produce ineffective self-explanations. Useful self-explanations can be categorized as either derivations or procedural explanations [12]. Derivations answer the question "Where did this step come from?" and procedural explanations answer the question "Why was this step done?"

When students do not engage in self-explanation, they do not tend to develop a deep understanding of the material. Novices tend to match surface features of a problem, like diagrams and problem statement wording, with those in a worked-out example. In contrast, experts use the principles and deep structure as criteria for matching a worked-out example to a problem [13]. The deep structure refers to a general plan or sequence of principle applications that can be followed in order to solve the problem. By providing worked-out examples with well-structured explicit steps to the solution and annotations of the relevant principles for each step, we are presenting students with examples of good self-explanations. This is expected to promote identification of the underlying problem structure and facilitate recognition of similar problem structure in different problems. Providing an annotated, worked-out example during problem solving enables a direct comparison and should encourage the student to focus on the common deep structure between the problem and the example. This will lead the students who are provided with these examples to perform better on tasks that test the deep structural understanding of the problems than those who are not provided with them.

2 Methodology

This was a hybrid study in that it was both naturalistic and experimental. The experiment was conducted during a second semester, college level physics course. As part of the graded homework for this course, students solved problems with Andes. Students who volunteered to participate in the experiment used a modified version of Andes to do this homework. The post-test for this study was administered either three or four days before the in-class exam, depending on students' regular lab sessions. The time frame of homework completion was at the students' discretion, and ranged from a few weeks before the relevant in-class exam to many weeks after. This unanticipated confound was resolved by the creation of a new category "No-Training" for participants who had not done any of their homework before this study's post-test.

The study had two experimental conditions: Examples and Hints. In the Examples condition, participants were presented with annotated, worked-out examples in response to any help request while using Andes. Each problem was mapped to a single example, but several problems were mapped to the same example if they shared the same deep structure. In the Hints condition, participants were given Andes' normal graded, step-dependent hints in response to help requests. The dependent variable for this experiment was performance on a problem matching task. Participants were asked to choose which of two problem statements would be solved most similarly to the given problem statement. This task is meant to evaluate deep learning by measuring participants' recognition of deep structure similarities.

The study participants were recruited from students already participating in the physics section of Pittsburgh Science of Learning Center LearnLab (`http://www.learnlab.org/`). The physics section was run as part of the General Physics I/II classes in the United States Naval Academy in Annapolis, Maryland. A total of forty-six volunteers were recruited from two sections of this course taught by the same professor. Participants were instructed to download the version of Andes which was modified for this study to use for the assigned homework problems on the topic of "Inductors." Because use of Andes was demonstrated in class and required for homework throughout the course, students in these sections were expected to be familiar with it. No restrictions were placed on use of the study-assigned Andes program, textbooks, professors, peers, or any other supplementary material. Due dates for Andes homework in the course were not rigidly enforced. Only the unmodified Andes version of the homework on Inductors was made available to the participants for in the Hints condition. Those in the Examples condition were assigned the same homework problems but instead were given access only to a modified version of Andes with the graded hints replaced with a worked-out example problem.

The worked-out examples were designed to be near-transfer problems where numeric values and some other surface features were changed. The solution to a homework problem requires solving for a variable in an equation while the worked-out example shows steps to solving a different variable in the same equation. For example, the equation for Ohm's law is $V = IR$ (voltage = current*resistance). If one homework problem gives values for V and R and asks the student to calculate I, and another gives values for V and I and asks for R, then the one worked-out example used for both of these questions would show steps for calculating V from given values for I and R. This relationship means that only five worked-out examples were needed for the ten homework problems. The problem solving steps in the examples were written and annotated with the principle used in each step, or with a list of the equations that were algebraically combined for a given step. The example was designed to show completed problem solving steps and solutions identical to those used in unmodified Andes problems. The principles in the annotations were linked to the appropriate Andes subject matter help pages so that the same body of problem-solving information was available to all participants.

The post-test was administered during the last lab session of the class prior to the in-class examination on this material. The test format was adapted from the similarity judgment task described by Dufresne, et. al [14]. It consisted of twenty multiple choice questions in random order, with randomly ordered answer choices, presented one at a time with thirty minutes given to complete the test. Each question contained three unsolved problems: a model problem and two comparison problems. Each of these problems consisted of a few sentences and a diagram. There were four possible types of relationship between the model problem and the two comparison problems:

I. Same surface features with different deep structure
II. Same surface features with the same deep structure
III. Different surface features with different deep structure
IV. Different surface features with the same deep structure

Only one of the comparison problems in each question had the same deep structure as the model problem (Type II and IV). The homework covered five different deep structure concepts. In the post-test, four questions were related to each deep structure concept, each with a different combination of surface feature relatedness. The theoretical strengths of this method of measuring competence include emphasis on deep structure and de-emphasis of algebraic skills [14]. The participants were given the following written instructions:

> "In the following evaluation, you will be presented with a series of problem statements. You do not have to solve the problems! Your task will be to read the first problem statement and then decided which of the following two problems would be solved most similarly to the first one."

In contrast to the format used by Dufresne, et. al [14] The model problems in this study were repeated as infrequently as possible (given the small number of variables in each equation). The present study also drew all correctly matched deep structure problems in the post-test from the assigned homework and worked-out examples.

The participants were assigned to the two experimental groups in a pairwise random fashion based on their cumulative Grade Point Average (GPA). This single criterion was used to balance the two groups in terms of previous performance without regard for other variables such as class section, gender, academic major, or age. Ten specific homework problems from the fourteen Inductance problems available in Andes were assigned to the class.

3 Results

Of the twenty-three participants assigned to each condition, only nine from the Examples condition and twelve from the Hints condition had asked for help from Andes to solve at least one of the homework problems before the study's post-test. Twenty other participants had not started working on the assignment and two of these did not complete the post-test either. Five participants had solved at least one homework problem in Andes without using the any of the available help. Only those participants who completed the post-test were included in the final analysis. Of those who did any of the homework, only those who asked for help at least once were exposed to the manipulated variable, and so the five participants who did not ask for help were excluded.

Variable Averages	**No-Training**	**Hints**	**Examples**	p
In-Class Circuit Exam	187 ± 21	178 ± 34	201 ± 38	0.5684
Total Training Time (s)	-	7942 ± 3681	4189 ± 2407	0.0735
Time per Problem (s)	-	672 ± 238	508 ± 162	0.2540
# of Problems Solved	-	11 ± 1.53	8.1 ± 2.8	**0.0427**
Weighted Post-Test	3.56 ± 0.49	4.30 ± 0.37	4.88 ± 0.56	0.0010
Problem Efficiency	-	0.413 ± 0.076	0.711 ± 0.215	**0.0034**

Fig. 1. Results are reported as mean±95% confidence limits

There were no significant differences in performance on circuit questions among the three conditions on the in-class examination administered before this study. This suggests that even though the participants who did not do homework self-selected the No-Training condition, all three conditions ended up with equivalently competent students. (see Fig. 1: In-Class Circuit Exam; $F_{(2,36)} = 0.57$, $p = 0.5684$). There was a notable but not statistically significant difference in the total time participants chose to spend solving homework problems between the Examples and Hints groups, with the Hints group spending more time on problems (see Fig. 1: Total Training Time; $t_{17.6\ corrected} = 1.90$, $p = 0.0735$). In contrast, the average time spent per problem (see Fig 1: Time per Problem; $t_{19} = 1.18$, $p = 0.2540$) was more consistent between the two groups. There was a significant difference in the average number of problems attempted between the two groups, with the Hints groups working on more problems than the Examples group (see Fig. 1: # of Problems Solved; $t_{19} = 2.17$, $p = 0.0427$).

By construction, the post-test questions varied considerably in their difficulty; for instance, it should be easier to identify similar deep structure when the surface features are also similar, and harder to identify them when the surface features are different. To more accurately measure competence, a weighted score was used. The post-test questions were weighted according to their difficulty as determined by the performance of the No-Training participants on each of the questions. The weight given to each question was $1 - \frac{\#correct}{18}$, where $\#correct$ was the number of participants from the No-Training condition who answered the given question correctly The calculated weights on the problems were in agreement with *a priori* expected performance differences on the different problem types.

When an ANOVA model was fit using the weighted post-test scores (see Fig. 1: Weighted Post-Test), a statistically significant difference among the three groups was detected ($F_{(2,36)} = 8.49$, $p = 0.0010$). With the Tukey-Kramer adjustment for multiple comparisons, it was found that the participants in the Examples condition did significantly better on the post-test than those in the No-Training condition ($t = 3.98$, $p = 0.0009$). The Hints condition also did better than the No-Training condition ($t = 2.45$, $p = 0.0496$). However, it was not possible to distinguish a difference between the Hints condition and the Examples condition based solely on the weighted post-test score ($t = 1.61$, $p = 0.2525$).

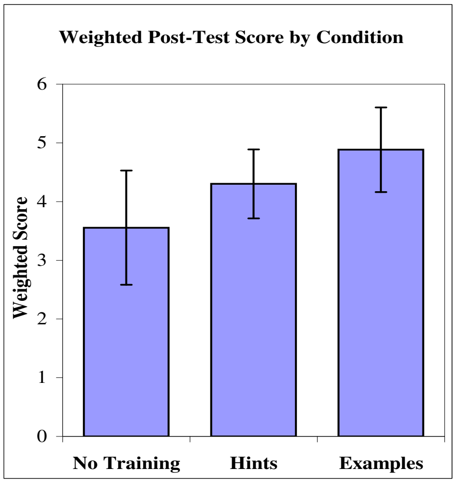

Fig. 2. Results are reported as means with error bars showing the ±95% confidence limits. Either form of training is better than none, but the difference in weighted post-test scores of the Hints and Examples conditions are not statistically significant.

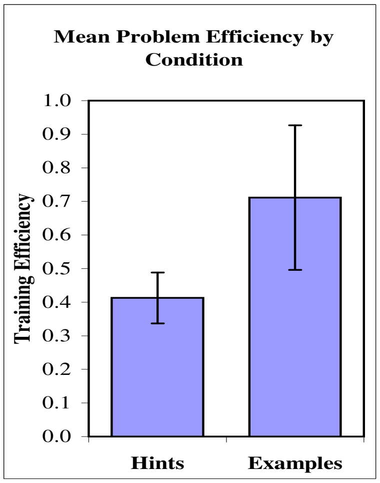

Fig. 3. Problem efficiency is defined as by condition where training efficiency is the weighted post-test score divided by the number of problems solved. Results are reported as means with error bars showing the ±95% confidence limits. Training problems with examples were more efficient at raising post-test scores than training problems with hints.

Other dependent variables, such as GPA and in-class examination scores, were not found to be significant factors in any ANCOVA models.

Problem efficiency was also calculated, that is, the increase in weighted post-test score per training problem done. The Examples and Hints conditions had weighted post-test scores that were not significantly different from each other but the participants in the Examples condition chose to do fewer training problems. The problem efficiency for the Examples condition was significantly higher than for the Hints condition (see Fig. 1:Problem Efficiency and 3; $t = 3.34$ $p = 0.0034$).

An ANCOVA model was fit using the weighted post-test scores with the number of problems attempted as the covariate and the Hints or Examples condition as the categorical variable (see Fig. 4. It was determined that the interaction effect between the condition and the number of problems was not significant ($p = 0.8290$). When the number of training problems was controlled by estimating the mean weighted post-test score at the overall mean number of training problems ($\mu_* = 9.76$), the difference in scores for the two training conditions condition was significant ($\mu_{Examples} = 4.88 \pm 0.46$; $\mu_{Hints} = 4.14 \pm 0.50$; $t = 2.30$, $p = 0.0338$). This was consistent with the Problem Efficiency results and demonstrates that given the same number of training problems, participants

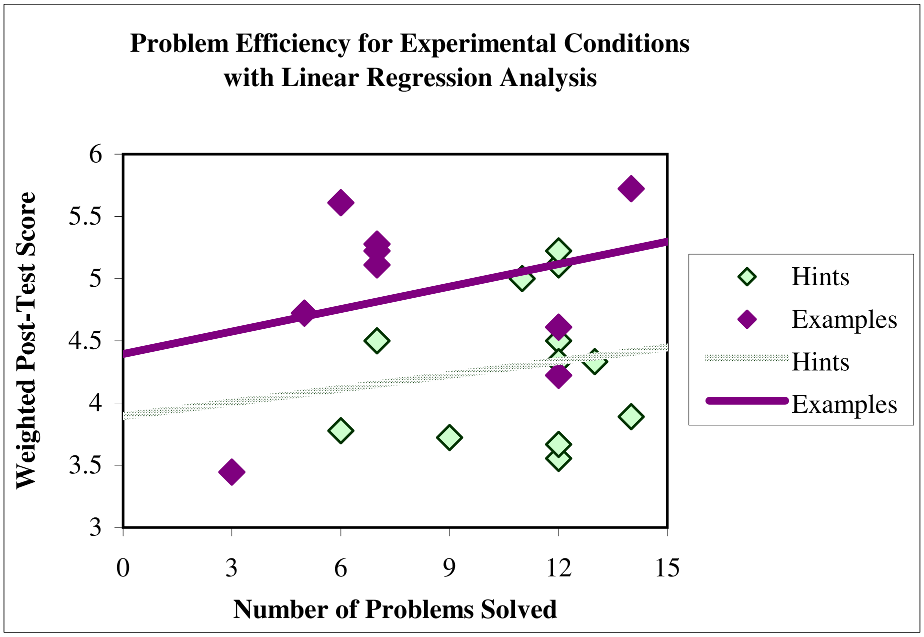

Fig. 4. Weighted post-test score versus number of problems solved with a fitted regression lines for the Hints and Examples conditions.

in the Examples condition performed better on the post-test than participants in the Hints condition.

4 Discussion

The results of this study demonstrate the value of working on Andes training problems to improve competence, whether with worked-out examples or graded hints. Although students in the No-Training condition were self-selected, they showed no significant difference in competence with basic circuit concepts prior to the study (as measured by scores on an in-class exam). One important difference between the two training conditions was the time on task, measured by the amount of time spent on the training homework. Participants in the Example condition chose to solve fewer training problems on average than the participants in the Hints condition. This was not due to the participants in the examples condition taking longer to solve problems, as the average time to solve each problem was not significantly different, but due to participants in the Hints condition choosing to spend more time working on more problems. Though participants in the Examples condition solved fewer problems on average than those in the Hints condition, they did at least as well on the post-test. This evidence supports the hypothesis that worked-out examples are a more efficient form of problem-solving help than graded hints. Due to the small number of participants involved in this study, aptitude treatment interactions could not be examined. A larger study might reveal an expertise reversal effect, where worked-out examples are more effective than graded hints for novices and less effective than graded hints for experts [15].

While previous studies have shown that providing worked-out examples can lead to shallow learning [13], this study indicates that worked-out examples may in fact be more efficient at promoting deep learning than graded hints. This has implications for tutoring system design in that examples may be a valuable addition to intelligent tutoring systems. Moreover, adding worked-out examples to an intelligent tutoring system should be fairly easy. The examples used in this study were easily added to Andes, mostly because there was no "intelligence" that needed to be designed to implement this strategy. One possible disadvantage of graded hint sequences is that they may be too rigid to accommodate the individual thought processes of different students. If the intelligent tutoring system that provides graded hints is good at assessing the participant's thought process, then the hints it can provide are likely to be effective. If the system can't identify and provide feedback relevant to a student's thought process, the hints will probably seem close to meaningless. If this happens too often, the student may decide that the hints are useless.

Worked-out examples can be a way of making sure that the system can provide a shared context for the student. They may be of particular value when the system has not collected enough data to evaluate the student effectively or if communication seems to have failed. One way to integrate worked-out examples into a graded hint system is to replace the bottom-out hint with the relevant worked-out example. This change may be particularly useful in addressing the help abuse [8]. It would be informative to see whether this strategy would reduce help abuse or provide incentive for the participants to click through a hint sequence rapidly just to see the worked-out example at the end. Worked-out examples could be integrated into more complicated intelligent tutoring systems that can assess the utility of different actions and provide these examples when appropriate [16]. Increasing the diversity of possible useful actions in such a system could only improve its performance.

References

1. Bloom, B.S.: The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. Educational Researcher **13** (1984) 4–16
2. Sweller, J., Cooper, G.A.: The use of worked examples as a substitute for problem solving in learning algebra. Cognition and Instruction **2**(1) (1985) 59–89
3. Cooper, G., Sweller, J.: Effects of schema acquisition and rule automation on mathematical problem-solving transfer. Journal of Educational Psychology **79**(4) (1987) 347–362
4. Brown, D.E.: Using examples and analogies to remediate misconceptions in physics: Factors influencing conceptual change. Journal of Research in Science Teaching **29**(1) (1992) 17–34
5. Catrambone, R.: Aiding subgoal learning - effects on transfer. Journal of Educational Psychology **87**(1) (1995) 5–17

6. Koehler, M.J.: Designing case-based hypermedia for developing understanding children's mathematical reasoning. Cognition and Instruction **20**(2) (2002) 151–195
7. VanLehn, K., Lynch, C., Schulze, K., Shapiro, J.A., Shelby, R., Taylor, L., Treacy, D., Weinstein, A., Wintersgill, M.: Andes physics tutoring system: Five years of evaluations. In Looi, G.M.C.K., ed.: Proceedings of the 12th International Conference on Artificial Intelligence in Education, Amsterdam, IOS Press (2005)
8. Aleven, V., Koedinger, K.R.: Investigations into help seeking and learning with a cognitive tutor. Papers of the AIED-2001 Workshop on Help Provision and Help Seeking in Interactive Learning Environments (2001) 47–58
9. LeFevre, J.A., Dixon, P.: Do written instructions need examples? Cognition and Instruction **3**(1) (1986) 1–30
10. Trafton, J.G., Reiser, B.J.: The contribution of studying examples and solving problems to skill acquisition. In: Proceedings of the 15th Annual Conference of the Cognitive Science Society, Hillsdale: Lawrence Erlbaum Associates, Inc. (1993) 1017–1022
11. Chi, M.T.H., Bassok, M., Lewis, M.W., Reimann, P., Glaser, R.: Self-explanations: How students study and use examples in learning to solve problems. Cognitive Science **13**(2) (1989) 145–182
12. Chi, M.T.H., VanLehn, K.A.: The content of physics self-explanations. Journal of the Learning Sciences **1**(1) (1991) 69–105
13. VanLehn, K., Johns, R.M.: Better learners use analogical problem solving sparingly. In Utgoff, P.E., ed.: Machine Learning: Proceedings of the Tenth Annual Conference, San Mateo, CA, Morgan Kaufmann Publishers (1993) 338–345
14. Dufresne, R.J., Gerace, W.J., Hardiman, P.T., Mestre, J.P.: Constraining novices to perform expertlike problem analyses: Effects on schema acquisition. Journal of the Learning Sciences **2**(3) (1992) 307–331
15. Kalyuga, S., Ayres, P., Chandler, P., Sweller, J.: The expertise reversal effect. Educational Psychologist **38**(1) (2003) 23–31
16. Murray, R.C., VanLehn, K.: Dt tutor: A decision-theoretic, dynamic approach for optimal selection of tutorial actions. In Gauthier, Frasson, VanLehn, eds.: Intelligent Tutoring Systems: 5th International Conference. Volume 1839 of Lecture Notes in Computer Science., Montreal, Canada, Berlin: Springer (2000) 153–162

A Appendix

– For examples of the training problems; the annotated, worked-out examples; and the post-test problems, visit `http://www.pitt.edu/~mringenb/AWOE/`.
– For more information about Andes, visit `http://www.andes.pitt.edu/`.
– For more information about LearnLab, visit `http://www.learnlab.org/`.