Preparing Students for Effective Explaining of Worked Examples in the Genetics Cognitive Tutor

Albert Corbett (corbett@cmu.edu)
Ben MacLaren (maclaren@andrew.cmu.edu)
Angela Wagner (awagner@cmu.edu)
Human-Computer Interaction Institute, Carnegie Mellon University
Pittsburgh, PA 15213 USA

Linda Kauffman (lk01@andrew.cmu.edu)
Aaron Mitchell (apm1@andrew.cmu.edu)
Department of Biological Sciences, Carnegie Mellon University
Pittsburgh, PA 15213 USA

Ryan S. J. d. Baker (rsbaker@wpi.edu)
Sujith M. Gowda (sujithmg@wpi.edu)
Department of Social Science and Policy Studies, Worcester Polytechnic Institute
Worcester, MA 01609 USA

Abstract

This study examines the impact of integrating worked examples into a Cognitive Tutor for genetics problem solving, and whether a genetics process modeling task can help prepare students for explaining worked examples and solving problems. Students participated in one of four conditions in which they engaged in either: (1) process modeling followed by interleaved worked examples and problem solving; (2) process modeling followed by problem solving without worked examples; (3) interleaved worked examples and problems without process modeling; or (4) problem solving alone. Tutor data analyses reveal that process modeling led to faster reasoning and greater accuracy in explaining problem solutions. Process modeling and worked examples together led to faster reasoning in problem solving than did any of the other three conditions. Students in all conditions achieved equivalent problem-solving knowledge, as measured by posttest accuracy, although the tutor results suggest reasoning speed may be a more sensitive measure of learning.

Keywords: Education; Problem solving; Learning; Intelligent Tutors; Worked Examples.

Introduction

It is well-documented that integrating worked examples with problem solving, either by interleaving full problem solutions with problems to be solved (Pashler, Bain, Bottge, Graesser, Koedinger, McDaniel & Metcalfe, 2007; Sweller & Cooper, 1985) or by gradually fading the number of solved steps that are provided (Renkl & Atkinson, 2003), serves to decrease total learning time and yields improved learning outcomes. Recently, several studies have examined the benefits of incorporating worked examples into intelligent tutoring systems (ITSs) for problem solving in a variety of domains: stoichiometry (Mclaren, Lim & Koedinger, 2008) algebra (Anthony, 2008; Corbett, Reed, Hoffman, MacLaren & Wagner, 2010b); geometry (Salden, Aleven, Schwonke & Renkl, 2010; Schwonke, Renkl,

Krieg, Wittwer, Aleven & Salden, 2009; Schwonke, Renkl, Salden & Aleven, 2011) and statistics (Weitz, Salden, Kim & Heffernan, 2010).

In these ITS studies, the chief benefit of incorporating worked examples has been to increase learning efficiency. The studies that report learning time universally find that interleaving worked examples (Corbett, et al, 2010b; McLaren, et al, 2008; Weitz, et al, 2010) or fading solution steps (Schwonke, et al, 2009) reduces learning time for a fixed set of activities compared to pure problem solving, primarily because students process worked solutions more rapidly than they can solve corresponding problems.

But unlike the classic worked-example literature, these ITS studies generally do not find that incorporating worked examples leads to more accurate posttest problem-solving than problem solving alone (Anthony, 2008; Corbett et al, 2010b; McLaren, et al, 2008; Schwonke, et al, 2009, 2011; Weitz, et al, 2010). The exception is Salden, et al (2010), who found that adaptively fading examples based on a model of each student's knowledge led to some relative improvement on posttest problem solving. Similarly, the evidence that students learn more deeply when worked examples are integrated into ITSs is mixed at best, although two papers report better retention of problem solving knowledge (Anthony, 2008; Salden, et al, 2010) and Schwonke, et al (2009) found evidence of greater conceptual transfer in one of two studies.

The present study examines the hypothesis that

- integrating worked examples and problem solving in an ITS will yield better learning outcomes when preceded by ITS learning activities that focus on domain knowledge relevant to the student explanations

This study examines worked examples and problem solving in the domain of genetics. The study employs an existing Cognitive Tutor for genetics problem solving, which has

been piloted at 15 universities around the country (Corbett, Kauffman, MacLaren, Wagner & Jones, 2010a). In this project we are developing two types of Cognitive Tutor (CT) activities to prepare students for deeper understanding in genetics problem solving: *self-explanation of worked examples* and *genetic process modeling*. In the worked example CT activities, students are given solved genetics problems and are asked to select menu-based explanations of each solution step, as in several earlier ITS studies (Salden, et al, 2010; Schwonke, et al, 2009, 2011; Weitz, et al, 2010). In *genetic process modeling* CT activities students reason directly about the underlying genetic processes that are relevant to a problem-solving task. This latter CT activity that focuses on developing domain knowledge prior to worked examples and problem solving is novel to the study presented here. In this study we examine student performance during learning across two sessions of CT activities, as well as on problem-solving posttests and measures of robust learning. The following section describes the problem-solving domain and CT activities.

The Domain: Three Factor Cross Gene Mapping

Genetics is a fundamental, unifying theme of biology and is viewed as a challenging topic by students and instructors, in part because it relies heavily on problem solving. This problem solving is characterized by abductive reasoning, in which students are given a set of observations and reason backwards to infer properties of the underlying genetics processes that produced the data. This study focuses on an abductive reasoning task that employs a gene mapping technique called a *three-factor cross* (3FC). In a 3FC problem, two organisms, e.g., fruit flies are crossed and the pattern of offspring phenotypes that result is analyzed to infer (1) the order of three genes that lie on one chromosome, and (2) the relative distances between gene pairs. Students can solve these problems algorithmically without reference to genetics, but the goal of this project is to ground student reasoning in the underlying process summarized in the following paragraph.

Figure 1a depicts the order of three genes on a chromosome pair belonging to a parent who is heterozygous for the genes. Ordinarily in reproduction, half this parent's offspring would inherit the three alleles on one chromosome (B, A, C), and half would inherit the other three alleles, (b, a, c). However, during meiosis, the two chromosomes in a homologous pair generally exchange genetic material. In some cases such a "crossover" will occur between two of the genes, A and B, as depicted in Figure 1b. As a result, some offspring will inherit the allele combination B, a, c
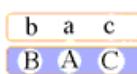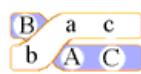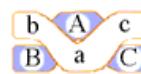

| Fig. 1a | Fig. 1b | Fig. 1c |

Figure 1: Three genes that appear on a chromosome pair in a parent organism (1a), and the impact of a single crossover (1b), or double crossover (1c) in meiosis.

from this parent and others will inherit the alleles b, A and C. A similar crossover can occur between the A and C genes, and very rarely, crossovers will occur between each gene pair as displayed in Figure 1c.

Figure 2 displays the Cognitive Tutor interface for a three-factor cross problem that results from a test cross involving the parent depicted in Figure 1. The table at the left of the screen represents the offspring that result from the cross. The letters represent observable traits, governed by the corresponding underlying genes, and because of crossovers in meiosis, all eight possible allele combinations are observed. Since the probability that a crossover occurs between two genes is proportional to the distance between them, the student reasons about the relative frequencies of the phenotypes to infer the middle gene and to calculate the distances among genes. In Figure 2 the student has almost finished the problem. To the right of the table, the student has summed the offspring in each of four phenotype groups and identified the type of each group (as parental, single crossover, or double crossover). The student has inferred the middle gene on the chromosome, and entered a gene sequence below the table. Finally, in the lower right the student has calculated the crossover frequency and distance between the middle gene and each of the two outer genes and will perform the last two steps for the two outer genes.
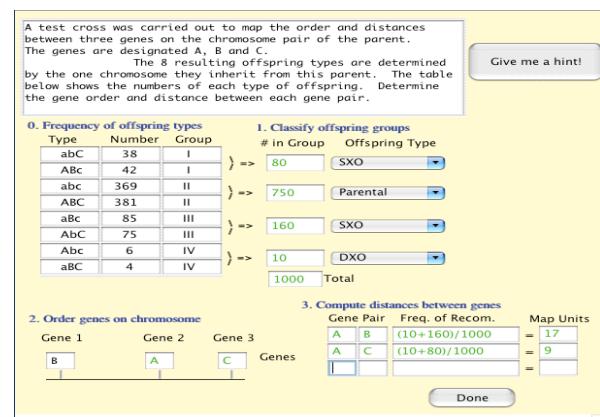

Figure 2: The CT interface for a 3FC problem-solving task.

Cognitive Model. The cognitive model for this task includes the following types of knowledge components (KCs), some of which apply more than once in a problem:

- Summing offspring numbers (2 KCs)
*Identifying offspring group type (3 KCs) (parental, single-crossover or double-crossover)
*Identifying the middle gene (1 KC)
*Calculating the frequency of crossovers (2 KCs) (for single crossovers and double crossovers)
- Calculating map unit distances (1 KC)

Formative analyses in this paper focus on the three starred types. The group-identification KCs hinge on offspring group size and are relatively easy for students. Identifying the middle gene requires the analysis of allele combinations

across groups and is a very challenging skill for students. The calculation KCs are also challenging, requiring students to identify the offspring groups relevant to each gene pair and to combine their respective frequencies arithmetically.

Worked examples. Figure 3 displays the interface for the 3FC worked-example activities. A complete 3FC problem solution is presented at the left of the screen. Students explain each step with the two menus to the right of the step. In the first menu, students select an explanation of the *empirical evidence* that led to the answer and in the second menu students select the *underlying genetic process* that explains the answer. As in all Cognitive Tutor activities, students receive accuracy feedback on each menu selection and can ask for help as needed for each menu.
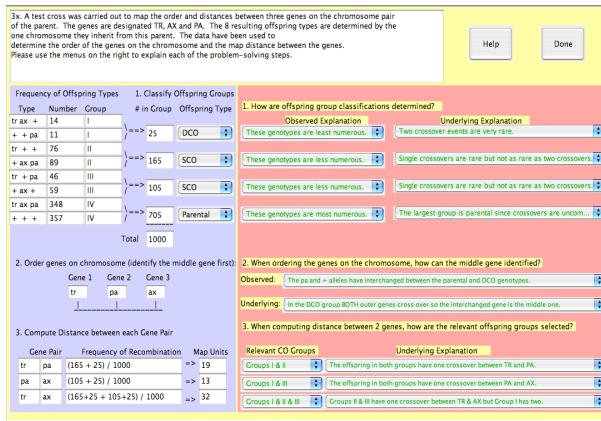

Figure 3: The CT worked example Interface

Genetic Process Modeling. Figure 4 displays the new CT process modeling activity in which students relate the underlying genetics processes and corresponding empirical data in a 3FC task. The table at the top of the screen has six main columns. Two columns at the left of the screen depict (unobservable) genetic crossovers in graphical and symbolic form. The next two columns to the right represent properties of the offspring that result from the crossovers. In each activity, students reason about the relationship among these observable and unobservable components of the process. The values for one of the four columns are given in each problem, and students generate the corresponding values for the other three columns. At the bottom of the screen, students select natural language summaries of the relationships from menus.

Method

Participants. Sixty-seven CMU undergraduates enrolled in either genetics or introductory biology courses were recruited to participate in this study for pay. Students were randomly assigned to one of four treatment groups.

Procedure. Students participated in two two-hour sessions on consecutive days and completed a problem-solving retention test one week later, as summarized in Table 1.
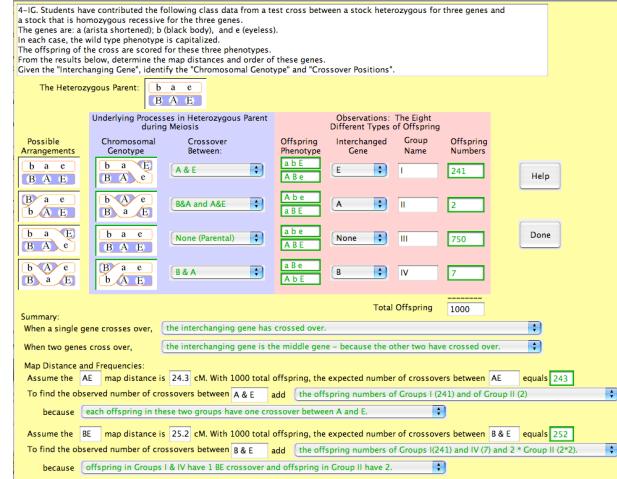

Figure 4: The CT interface for the process modeling task.

In Session 1, students completed conceptual knowledge and problem solving pretests, then completed Cognitive Tutor learning activities, followed by a conceptual knowledge posttest. In Session 2, students completed Cognitive Tutor problems, followed by a problem solving posttest, a transfer test, and a preparation for future learning (PFL) test.

Table 1: Student activities in the three study sessions.

Session 1: Pretests: Conceptual Knowledge & Problem Solving Cognitive Tutor Activities (Four Conditions) Conceptual Knowledge Posttest
Session 2: Cognitive Tutor Problem Solving Posttests: Problem Solving, Transfer & PFL
Session 3: One-week Retention test: Problem Solving

Design. There were four conditions in the study, defined by CT learning activities in the first session:

- **Process Modeling (MOD):** Students completed process modeling activities for up to 30 minutes, followed by up to 30 minutes of problem solving. (N=16 students)
- **Interleaved Worked Examples (IWE):** Students completed interleaved worked examples and problem solving activities for up to 60 minutes. (N=20)
- **Process Modeling and Interleaved Worked Examples (ALL):** Students completed process modeling activities for up to 20 minutes, then interleaved worked examples and problem solving for up to 40 minutes. (N=18)
- **Problem Solving (PS):** Students exclusively completed problem solving activities. (N= 13)

In Session 2 all students completed problem-solving activities for up to 60 minutes.

Tests. We developed four types of tests for the study:
- **Problem Solving Tests**: Three forms were developed. Within each condition, each form served as a pretest for 1/3 of the students, a posttest for 1/3 of the students and a one-week retention test for 1/3 of the students.

- **Conceptual Knowledge Tests**: Two forms were developed of this multiple choice test that tapped students' understanding of crossovers in meiosis. Each form served as a session-1 pretest for half the students and a session-1 posttest for the other half.
- **Transfer Tests:** A transfer test with two problems was administered following all CT activities. The first was a three-factor cross problem that required students to improvise an alternative solution. The second problem asked students to extend their reasoning to four genes.
- **Preparation for Future Learning (PFL):** This test presented a 2.5-page description of the reasoning in a four-factor cross experiment, then asked students to solve a four-factor cross problem.

Results

Average scores on the CK pretest were quite high, 86% correct, while average scores on the PS pretest were low, 34% correct. In two ANOVAs, there were no reliable differences among the four groups on either pretest.

Session 1 Cognitive Tutor Activities

The average number of cognitive tutor activities completed by students in the four conditions is displayed in Table 2. As can be seen, students in the baseline problem-solving condition completed an average of almost two more abductive problems than students in the other conditions. However, some students in this condition completed the full set of problems in less than an hour, so total time on task in session 1 was somewhat lower in this condition.

Table 2: Mean number of Day-1 Cognitive Tutor activities.

Condition	Process Modeling	Worked Examples	Problem Solving	Time (min.)
MOD	3.4	---	4.4	44.6
IWE	---	5.4	5.2	45.0
ALL	2.4	5.1	5.2	49.3
PS	---	---	6.8	40.3

Worked-Example Tasks To begin examining the impact of process modeling on learning, we compared students' performance on the worked example tasks for the ALL and IWE conditions. Table 3 displays both average student accuracy and average time to explain three types of solution components in the 3FC worked example activities. For each step, students both describe the relevant observable data and explain the underlying genetic process.

As can be seen students in the ALL group who completed process modeling activities before the worked examples are 40% faster in explaining solution steps than students in the IWE group. In an ANOVA, this main effect of condition is reliable, $F(1,36)=21.16$, $p < .01$. The main effect of knowledge component (KC) type is reliable $F(2,72)=26.75$, $p < .01$ as is the main effect of explaining the observable data vs. underlying process, $F(1,36)=82.87$, $p < .01$. In pairwise t-tests, students in the ALL group are reliably

Table 3: Average accuracy (percent correct) and time (seconds) to explain three categories of observed actions and the underlying genetics in the CT worked examples.

	ALL Acc. %C	IWE Acc. %C	ALL Time sec.	IWE Time sec.
Identify Offspring Classes				
Describe Observation	92	93	10 [a]	19[a]
Explain Genetic Process	84	87	12 [a]	17 [a]
Identify Middle Gene				
Describe Observation	56	62	33 [a]	53 [a]
Explain Genetic Process	96	80	8 [a]	14 [a]
Calculate Crossover Frequency				
Describe Observation	82	72	18	22
Explain Genetic Process	75	63	10 [a]	27 [a]

[a] t-test is reliable, p < .01

faster than the IWE students for 5 of the 6 types of explanations, as shown in Table 3.

Students in the ALL group are only about 6% more accurate than the IWE students. In an ANOVA, the main effect of condition is not reliable, $F(1,36)=0.90$, while the main effect of knowledge component (KC) type is reliable $F(2,72)=16.29$, $p < .01$, as is the main effect of explaining the observable data vs. underlying process, $F(1,36)=4.19$, $p < .05$. One interaction, of explanation type and KC type, is reliable, $F(2,72)=25.00$, $p < .01$. Students were more accurate in explaining the observable data than the underlying process for offspring class identification and crossover frequency, but more accurate at explaining the process than the data for finding the middle gene.

Problem Solving In session 1, the first three problems that students in each group solved were identical. Table 4 displays the average accuracy of the four groups for the three types of solution steps. As can be seen, students in the two groups that completed a worked example prior to each problem are more accurate in completing the problem steps: they are about 11% more accurate in identifying offspring classes, 11% more accurate in calculating distances between gene pairs, and 20% more accurate in identifying the middle gene. In the accuracy data, there is little evidence that completing the process modeling activities improved problem-solving accuracy; accuracy levels are similar in the ALL and IWE conditions, and they are similar in the MOD and PS conditions. In an ANOVA, the main effect of condition is marginally reliable, $F(3,63)=2.22$, $p < .10$). The main effect of knowledge component type is reliable, $F(2,126)=20.69$, $p < .01$, but the interaction is not reliable.

Table 4 also displays the average time taken to complete the three types of problem solving actions. Across the 3 types of activities, students in the ALL condition completed the problem solving actions much faster than students in the other conditions. Across the three types of steps, IWE students took 26% more time, MOD students took 62%

Table 4: Average accuracy (percent correct) and time (seconds) to complete session-1 problem solving actions.

	MOD	IWE	ALL	PS
	Acc. %C	Acc. %C	Acc. %C	Acc. %C
Identify Offspring Classes	80	86	87	76
Identify Middle Gene	58	72	72	62
Calculate Crossover Frequency	60	60	70	57
	Time sec.	Time sec.	Time sec.	Time sec.
Identify Offspring Classes	4.0	4.2	2.7	8.1
Identify Middle Gene	40.7	32.7	24.2	56.8
Calculate Crossover Frequency	69.1	51.5	43.2	86.5

Table 5: Average accuracy (percent correct) and time (seconds) to complete session-2 problem solving actions.

	MOD	IWE	ALL	PS
	Acc. %C	Acc. %C	Acc. %C	Acc. %C
Identify Offspring Classes	91	96	96	95
Identify Middle Gene	61	72	82	80
Calculate Crossover Frequency	71	80	88	90
	Time sec.	Time sec.	Time sec.	Time sec.
Identify Offspring Classes	2.2	2.1	1.9	2.9
Identify Middle Gene	26.3	23.6	15.4	23.3
Calculate Crossover Frequency	24.6	21.9	15.3	18.9

more time and baseline PS students took 116% more time. In an ANOVA, the main effect of condition is reliable, $F_{(3,63)}=9.51$, $p < .01$. The main effect of knowledge component type is also reliable, $F_{(2,126)}=123.23$, $p < .01$, and the interaction is reliable, $F_{(6,126)}=2.68$, $p < .05$.

Session 2 Cognitive Tutor Problem Solving

Student performance in the second session, in which all students worked on the same set of 3FC CT problems serves as a measure of Session-1 learning outcomes. Figure 5 displays average time to solve the first five CT problems – the problems finished by all the students. As can be seen, students in the ALL group finished the problems more quickly than students in the other three groups. In a two-way ANOVA, the main effect of condition is marginally significant $F_{(3,63)} = 2.62$, $p < .06$. The main effect of problem number is also significant, $F_{(4,252)}=21.40$, $p < .01$, while the interaction is non-significant.

Table 5 displays the average accuracy of the four groups for the three types of solution KCs in these five problems. By the second session, there is little difference in accuracy among the ALL, IWE and PS groups, but the MOD group who engaged in processing modeling without worked examples lags behind the other groups. In an ANOVA, the main effect of condition is reliable, $F_{(3, 63)}=4.48$, $p < .01$.
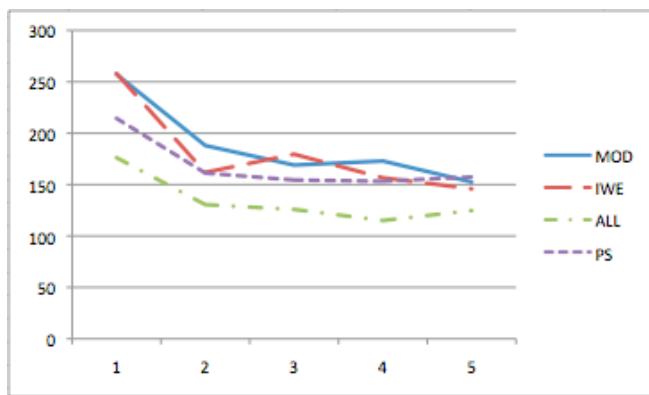

Figure 5: Average time to complete the first five 3FC CT problems in Session 2.

In t-tests the differences are reliable between the MOD and ALL groups and between the MOD and PS groups, and marginally reliable ($p < .06$) between the MOD and IWE groups. No pairwise differences among the ALL, IWE and PS groups are reliable. In the ANOVA, the main effect of KC type is also reliable, $F_{(2,126)}=29.03$, $p < .01$. The difference between the MOD group and the other groups appears larger for the harder KCs, but the interaction of KC and condition is not reliable.

Table 5 also displays the average time taken to complete the three types of problem solving actions. As can be seen, students in the ALL condition appear to reason about the problems more quickly than students in the other conditions. Across the three component types, students in the PS condition are 38% slower, students in the IWE condition are 46% slower and students in the MOD condition are 63% slower. Surprisingly, in an ANOVA, neither the main effect of condition nor the interaction of KC and condition is statistically significant. Inspecting the data set reveals that the variance in the ALL condition, 41.4, is much lower than the variance of the other conditions (PS = 279, IWE = 431, MOD = 977). While the values in the lower half of the four distributions are very similar (i.e., the faster students look similar in all four conditions), the PS, IWE and MOD distributions have longer tails at the high end, hinting at another interaction in these Day 2 KC times: The ALL condition appears to be especially helpful for the less prepared students, reducing their reasoning times.

Table 6 displays average accuracy for the conceptual knowledge pretests and posttests administered before and after the Day-1 CT activities. Average pretest scores across the four groups are already quite high, about 86% correct, and are almost unchanged on the posttest, averaging 87% correct. In a two-way ANOVA, the main effects of test and of condition, and the interaction are all non-significant.

Table 6 also displays the problem solving pretests given before the Day-2 CT activities and problem solving posttests given immediately following Day-2 CT activities. There are large learning gains across the four groups, with pretest accuracy averaging 34% correct and posttest scores averaging 84% correct. In a two-way ANOVA, the main

effect of test is reliable, F(1,63)=8.39, p < .01, but the main effect of condition and the interaction are not significant.

Table 6: Average accuracy (percent correct) on the conceptual knowledge and problem solving pretests and posttests and the robust learning tests in the four conditions.

	MOD	IWE	ALL	PS
	Acc. %C	Acc. %C	Acc. %C	Acc. %C
Conceptual Knowledge Pretest	81	89	92	83
Conceptual Knowledge Posttest	84	85	92	87
Problem Solving Pretest	29	36	33	37
Problem Solving Posttest	79	86	87	85
Transfer Test	88	85	85	88
Preparation for Future Learning	91	89	90	92
Problem Solving Retention	73	85	76	91

Table 6 also displays average student accuracy on the three measures of robust learning: the transfer test and PFL test administered at the end of Session 2 and the problem solving retention test administered a week later. As can be seen, average performance on the transfer and PFL tests is quite high across conditions and in two one-way ANOVAs, the main effect of condition was non-significant for both tests. On the delayed problem solving test, the PS baseline condition shows some sign of better retention, but the main effect of condition is not significant in an ANOVA.

Discussion and Conclusion

Several conclusions emerge from the results. Students in the ALL condition, who actively reviewed the underlying genetics processes before explaining interleaved worked examples, demonstrated the greatest learning throughout the first two days – as reflected in reasoning speed more strongly than in accuracy. There is one caveat, that while students in the baseline PS condition completed more problem-solving activities than the other three groups, they spent less total time on task the first day. Process modeling alone tended to yield lower problem-solving accuracy across both sessions, suggesting that reasoning about the underlying genetics processes alone is not especially useful for learning to solve problems, unless worked examples are provided to scaffold the relationship between the genetics processes and problem solving. As found in earlier studies integrating worked examples with intelligent learning environments, the IWE worked-example alone condition yielded problem solving performance similar to the baseline PS condition.

As in most of the prior ITS/worked-example studies, no differences were found among the conditions in accuracy gains on the problem-solving tests. Nor were differences found on the tests of robust learning. However, the analysis of CT performance implies that, while test performance accuracy may be of greater practical relevance, posttest performance time measures may be necessary to detect genuine differences in learning outcomes.

References

Anthony, L. (2008). *Developing handwriting-based intelligent tutors to enhance mathematics learning.* Unpub. doctoral dissertation, Carnegie Mellon University.

Corbett, A., Kauffman, L., MacLaren, B., Wagner, A., & Jones, E. (2010a). A Cognitive Tutor for genetics problem solving: Learning gains and student modeling. *Journal of Educational Computing Research*, *42*, 219-239.

Corbett, A.T., Reed, S.K., Hoffmann, R., MacLaren, B., Wagner, A. (2010b). Interleaving worked examples and cognitive tutor support for algebraic modeling of problem situations. In *Proceedings of the Thirty-Second Annual Meeting of the Cognitive Science Society* (pp. 2882-2887).

McLaren, B. M., Lim, S.-J., & Koedinger, K. R. (2008). When is assistance helpful to learning? Results in combining worked examples and intelligent tutoring. In *Proceedings of the 9th International Conference on Intelligent Tutoring Systems* (pp. 677-680).

Pashler, H., Bain, P., Bottge, B., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). *Organizing Instruction and Study to Improve Student Learning*. Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.

Renkl, A. & Atkinson, R. K. (2003). Structuring the transition from example study to problem solving in cognitive skills acquisition: A cognitive load perspective. *Educational Psychologist, 38*, 15-22.

Salden, R. J. C. M., Aleven, V., Schwonke, R., & Renkl, A. (2010). The expertise reversal effect and worked examples in tutored problem solving. *Instructional Science*, 38, 289-307.

Schwonke, R., Renkl, A., Salden, R., & Aleven, V. (2011). Effects of different ratios of worked solution steps and problem solving opportunities on cognitive load and learning outcomes. *Computers in Human Behavior*, 27, 58-62.

Schwonke, R., Renkl, A., Krieg, C., Wittwer, J., Aleven, V., & Salden, R. (2009). The worked-example effect: Not an artifact of lousy control conditions. *Computers in Human Behavior, 25*, 258-266.

Sweller, J. & Cooper, G.A. (1985). The use of worked examples as a substitute for problem solving in learning algebra, *Cognition and Instruction, 2*, 59–89.

Weitz, R., Salden, R. J. C. M., Kim, R. S., & Heffernan, N. T. (2010). Comparing worked examples and tutored problem solving: Pure vs. mixed approaches. In *Proceedings of the Thirty-Second Annual Meeting of the Cognitive Science Society* (pp. 2876-2881).