Adapting to When Students Game an Intelligent Tutoring System

Ryan S.J.d. Baker[1], Albert T. Corbett[2], Kenneth R. Koedinger[2],
Shelley Evenson[3], Ido Roll[2], Angela Z. Wagner[2], Meghan Naim[4],
Jay Raspat[4], Daniel J. Baker[5], and Joseph E. Beck[6]

[1] Learning Sciences Research Institute, University of Nottingham, Nottingham, UK
Ryan.Baker@nottingham.ac.uk
[2] Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA, USA
{corbett, koedinger, iroll, awagner}@cmu.edu
[3] School of Design, Carnegie Mellon University, Pittsburgh, PA, USA
evenson@andrew.cmu.edu
[4] North Hills Junior High, Pittsburgh, PA, USA
{raspatj, naimm}@nhsd.k12.pa.us
[5] Department of Pediatrics, University of Medicine and Dentistry of New Jersey,
New Brunswick, NJ, USA
bakerd1@umdnj.edu
[6] Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA
joseph.beck@gmail.com

Abstract. It has been found in recent years that many students who use intelligent tutoring systems game the system, attempting to succeed in the educational environment by exploiting properties of the system rather than by learning the material and trying to use that knowledge to answer correctly. In this paper, we introduce a system which gives a gaming student supplementary exercises focused on exactly the material the student bypassed by gaming, and which also expresses negative emotion to gaming students through an animated agent. Students using this system engage in less gaming, and students who receive many supplemental exercises have considerably better learning than is associated with gaming in the control condition or prior studies.

1 Introduction

In recent years, increasing attention has been paid to the subject of how students choose to use intelligent tutoring systems. Recent models have suggested that students adopt a variety of strategies for using intelligent tutoring systems and other interactive learning environments, with different strategies potentially leading to different learning outcomes [2,3,7,14]. One strategy in particular, gaming the system, has been found to be associated with poorer learning gains in intelligent tutoring systems [5,7]. We define gaming the system as attempting to succeed in an educational environment by exploiting properties of the system rather than by learning the material and trying to use that knowledge to answer correctly. Gaming has been observed in a variety of types of learning environments, from educational games [10] to online newsgroups [9], and has been repeatedly documented in intelligent tutoring systems [1,7,8,12,13].

M. Ikeda, K. Ashley, and T.-W. Chan (Eds.): ITS 2006, LNCS 4053, pp. 392 – 401, 2006.

Within the specific intelligent tutoring system that we will discuss in this paper, gaming behavior consists of systematic guessing and rapid-fire hint requests [4].

Baker and his colleagues [4] have determined that gaming can be divided in some systems into two distinct behaviors – "harmful" gaming, which typically occurs on the problem steps the student knows least well, and is associated with poor learning outcomes, and "non-harmful" gaming, which typically occurs on problem steps the student already knows, and is not associated with poor learning outcomes.

In this paper, we present a tutor component that responds to harmful gaming, in order to improve gaming students' learning. This tutor incorporates an animated agent, Scooter the Tutor, who observes students as they interact with the tutor, looks increasingly unhappy when students game and gives a student supplementary exercises on the exact steps of the problem-solving process that the student gamed.

2 Design

Two previous attempts to address gaming in intelligent tutoring systems took a "preventative" approach to addressing gaming, attempting to directly prevent known gaming behaviors [1,8]. Researchers at Carnegie Mellon and Carnegie Learning introduced a two-second delay between each level of a multi-level hint, to prevent a student from clicking through hints at high speed, and gave mandatory hints ("proactive help") when a student commits more than three errors on a single step, preventing systematic guessing [1]. Researchers at the University of Massachusetts re-designed their system to not give help until a student had spent a minimum amount of time on the current problem [8].

In [7], we hypothesized that students using a system re-designed to directly prevent gaming would attempt to discover new ways to game. Shortly after, [13] found that students using a tutor with two-second help delays developed new strategies for gaming – for example, rapidly repeating the same error several times in a row in order to elicit delay-free proactive help. An additional concern with direct prevention is that students game features which are used in more positive ways by the majority of students who do not game.

Our design approach, by contrast, attempted to meet two conditions: First, the design must improve the learning of students who currently game. Second, the design must change the tutor minimally for students who do not game.

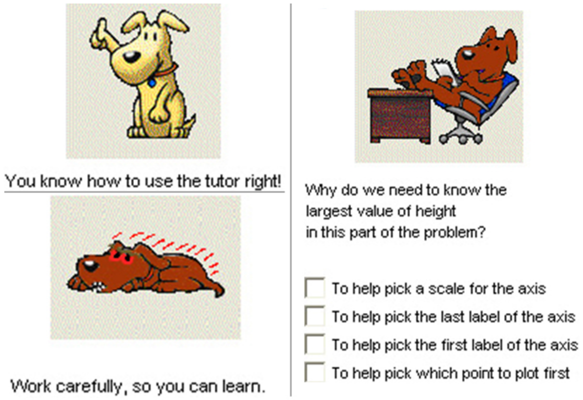
In accordance with these design goals, we developed a new component for the students' intelligent tutoring software – an animated agent named "Scooter the Tutor", developed using graphics from the Microsoft Office Assistant [11] but modifying those graphics to enable a wider range of emotions. Scooter was designed to both reduce the incentive to game, and to help students learn the material that they were avoiding by gaming, while affecting non-gaming students as minimally as possible.

When the student is not gaming, Scooter looks happy and occasionally gives the student positive messages (see the top-left of Figure 1). Scooter's behavior changes when the student is detected to be gaming harmfully (using an updated version of the gaming detector presented in [4,6]). If the detector assesses that the student has been gaming harmfully, but the student has not yet obtained the answer, Scooter displays increasing levels of displeasure (culminating in the expression shown on the bottom-left of Figure 1), to signal to the student that he or she should now stop gaming, and try to get the answer in a more appropriate fashion.

If the student obtains a correct answer through gaming, Scooter gives the student a set of supplementary exercises designed to give the student another chance to cover the material that the student bypassed by gaming this step. The supplementary exercises have three levels, each multiple-choice – the student is only given one chance to answer each level. In each of the first two levels of an exercise, the student is asked to answer a question that either requires understanding one of the concepts required to answer the step the student gamed through, or a question which is about what role the step they gamed through plays in the overall problem-solving process. If the student gets both the first and second levels wrong, he or she is given a third level, which is still relevant to the step the student gamed through, but which is very easy, in order to prevent indefinite floundering.

If the student gets any level right on the first try, Scooter lets the student return to the regular tutor exercise; if the student gets all three levels (including the very easy third level) wrong, Scooter assumes that the student was trying to game him, asks the student to attempt to get his exercises correct on the first try, and marks the problem step involved to receive supplementary exercises in future problems. If the student tries to game a supplementary exercise, Scooter displays anger.

Our goal, in designing Scooter, was to benefit students in three fashions. First, by representing how much each student had been gaming, Scooter both serves as acontinual reminder that the student should not game, and lets teachers know which students were gaming recently. Second, Scooter was intended to invoke social norms in students by expressing negative emotion when students game. Scooter's display of anger is a natural social behavior in this context; if a student systematically guessed every number from 1 to 38 when working with a human tutor, it seems reasonable to



You know how to use the tutor right!

Work carefully, so you can learn.

Why do we need to know the
largest value of height
in this part of the problem?

☐ To help pick a scale for the axis
☐ To help pick the last label of the axis
☐ To help pick the first label of the axis
☐ To help pick which point to plot first

Fig. 1. Scooter the Tutor – looking happy when the student has not been gaming harmfully (top-left), giving a supplementary exercise to a gaming student (right), and looking angry when the student is believed to have been gaming heavily, or attempted to game Scooter during a supplementary exercise (bottom-left)

expect that the human tutor would become impatient or upset. Therefore, we hypothesized that when Scooter becomes angry, he will invoke social norms, leading the student to game the system less. Third, by giving students supplemental exercises targeted to the material the student was gaming through, Scooter gives students a second chance and another way to learn material he or she may otherwise miss entirely. Additionally, supplemental exercises may change the incentive to game – whereas gaming might previously have been seen as a way to avoid work, it now leads to extra work. Thus, we predicted that Scooter would both reduce gaming and improve gaming students' learning, either by reducing their gaming or giving them a second chance to learn the material they miss by gaming.

3 Study Methods

We studied Scooter's effectiveness in the context of a year-long Cognitive Tutor curriculum for middle school mathematics, within 5 classes at 2 schools in the Pittsburgh suburbs. The study was conducted in the spring semester, after students had already used the Cognitive Tutor for several months.

Initially, the study was designed such that every student used both a version of the tutor with Scooter (experimental condition), and a version of the tutor without Scooter (control condition). Each student was randomly assigned to use one of two lessons (a lesson on percents, and a lesson on scatterplots) with Scooter, and the other lesson without Scooter. All students completed the control condition of the study first, and the experimental condition second. However, due to a scheduling error, the experimental condition of the study took place in the same week as subject material on percents was being taught in class. To avoid bias in favor of the experimental condition, we will therefore limit our discussion to data from the scatterplot lesson. 51 students participated in the experimental condition for the scatterplot lesson (12 were absent for either the pre-test or post-test, and thus their data will not be included in analyses relevant to learning gains); 51 students participated in the control condition for the scatterplot lesson (17 were absent for either the pre-test or post-test).

Before using the tutor, all students first viewed conceptual instruction, delivered via a PowerPoint presentation with voiceover and simple animations [cf. 4]. In the experimental condition, a brief description of Scooter was incorporated into the instruction. Then students completed a pre-test, used the tutor lesson for 80 minutes across multiple class periods, and completed a post-test. Test items were counterbalanced across the pre-test and post-test, and were identical to items used in past studies using this tutor lesson [4]. Log files were used to distill measures of Scooter's interactions with each student, including the frequency with which Scooter got angry, and the frequency with which Scooter gave a student supplementary exercises. In addition, observational data was collected to determine each student's frequency of gaming, using the quantitative observational method as in [7], in order to analyze Scooter's effects on gaming frequency. Another potential measure, the gaming detector [4], was not used because of risk of bias in using the same metric both to drive interventions and as a measure of the intervention's effectiveness.

4 Results

Scooter was associated with a sizeable, though only marginally significant, reduction in the frequency of observed gaming. 33% of students were seen gaming in the control condition, while 18% of students were seen gaming in the experimental condition, a marginally significant difference, $\chi^2(1,N=102)= 3.30$, p=0.07. However, although fewer students gamed, those students who did game did not appear to game less. The average gamer in the control condition gamed 17% of the time, while the average gamer in the experimental condition gamed 14% of the time, which was not a significant difference, t(23)=0.74, p=0.47.

Despite the apparent reduction in gaming, however, there was not an overall improvement in learning. Overall, students in the control condition averaged a 22 point pre-post gain (44%->66%), while students in the experimental condition averaged a 25 point pre-post gain (37%->62%), which was not a significant difference, t(70)=0.34, p=0.73. However, analyzing overall learning may not be the most appropriate way to test the intervention's effect on learning. Gamers are a fairly small subset of the overall population, both in this study and past studies [cf. 6,7].

Therefore, differences in gamers' learning may be swamped by normal variation in the rest of the population. Additionally, since students engaged in different degrees of gaming, and the detector was accurate but not perfect [cf.4], not all students who in engaged in harmful gaming received the same number of interventions from Scooter. Thus, in the following sections, we will look at the students who got a considerable amount of each type of intervention from Scooter, to see if and how the students' behavior and learning was affected by Scooter. We will analyze the two types of interventions separately, since the two types of interventions were given in different situations and may have had different effects.

4.1 Supplementary Exercises

Overall, Scooter gave a fairly small number of exercises: no student received a set of exercises from Scooter on more than 3.2% of problem steps (12 sets), and the median student received a set of exercises on only 1.1% of problem steps (3 sets). However,

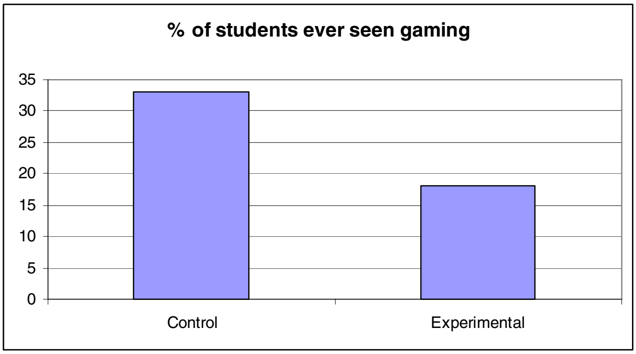


Fig. 2. The frequency of gaming (observed) in each condition

Scooter's exercises were assigned to exactly the problem steps students gamed on (according to the detector), and were significantly correlated to the frequency of observed gaming, r=0.43, F(1,38)=8.24, p<0.01, so the exercises might have had more effect on learning than their low frequency might otherwise indicate.

One possible model for how learning could relate to the number of supplementary exercises received is a linear relationship – the more supplementary exercises a student receives, the more they learn. However, students who never receive supplementary exercises don't receive supplementary exercises precisely because they don't engage in harmful gaming, and not engaging in harmful gaming is generally associated with better learning [cf. 4]. Therefore, if supplementary exercises positively affect learning, it may be more reasonable to expect students who receive either many or very few supplementary exercises to show good learning, with the students in the middle showing poorer learning.

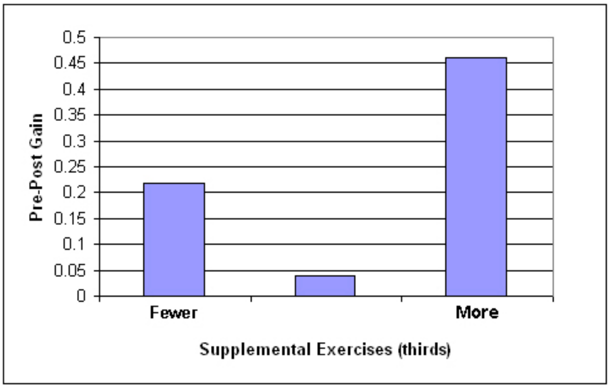


Fig. 3. The Learning Gains Associated With Receiving Different Levels of Supplemental Exercises From Scooter

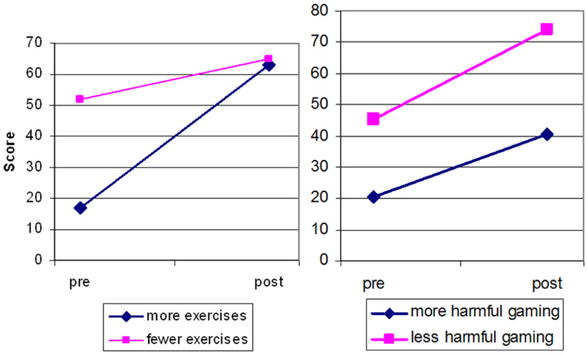


Fig. 4. Left: The Learning Gains Associated With Receiving Different Levels of Supplemental Exercises From Scooter (Top Third versus Other Two Thirds)**.** Right: The Learning Gains Associated With Different Levels of Harmful Gaming, in the Control Condition (Top Half of Harmful Gaming Versus Other Students).

In fact, this is exactly the relationship we find, as shown in Figure 3. The third of students that received the most supplementary exercises had significantly better learning than the other two thirds, t(37)=2.25, p=0.03; the overall difference between all three groups was also significant, F(2,36)=3.10, p=0.06.

This occurred because the students who received the most supplementary exercises started out behind the rest of the class (common among students who frequently game [cf. 7]), but caught up by the post-test (see Figure 4 Left). There was a statistically significant interaction between pre-test and post-test scores, and how many supplementary exercises the student received (top third versus other two thirds), F(1,37) = 5.07, p=0.03, for a repeated measures ANOVA. Note that there was not a ceiling in the mid-60s, nor a post-test floor effect: students in each group had perfect post-test scores, or low post-test scores.

In considering the evidence that students who received many supplemental exercises caught up to the rest of the class, it is worth remembering that students receive supplemental exercises because they are detected to be engaging in a large amount of harmful gaming. In both the control condition (see Figure 4 Right), and in prior studies involving the same tutor lesson [4,5], frequent harmful gaming is associated with starting out lower than the rest of the class, and falling further behind by the post-test, rather than catching up. As shown in Table 1, students in the control condition and past studies who did not use Scooter and engaged in more than the median amount of harmful gaming (among harmful gamers) averaged a 22 point learning gain, less than half of the average learning gain (46 points) of students who received many supplementary exercises, a statistically significant difference, t(47)=2.09, p=0.04.

Table 1. Learning gains for students who received large numbers of supplementary exercises from Scooter, and for students who did not use Scooter and engaged in more than the median amount of harmful gaming, among harmful gamers. All students used the same lesson on Scatterplots.

Group	Learning Gain
Experimental condition: more supplementary exercises	46 points
Control condition: more harmful gaming 2004: more harmful gaming [e.g. 5] 2003: more harmful gaming [e.g. 7]	20 points 18 points 25 points

Interestingly, although Scooter's exercises appear to be associated with improved learning, Scooter's exercises were not directly associated with the decrease in gaming reported in the previous section. If receiving an exercise from Scooter led a student to reduce his/her gaming, we would expect the students who received more exercises to reduce their gaming over time. There is no evidence of such a decrease. Figure 5 (left) shows the frequency in gaming over the 3 days of the study among the students who received many exercises (top third) in the experimental condition, compared to the other students. Among the students who received more exercises, neither the apparent increase in gaming from day 1 to day 2, nor the apparent decrease in gaming from day 2 to day 3, was statistically significant, $\chi^2(1,N=155)= 0.31$, $p=0.58$, $\chi^2(1,N=105)= 0.17$, $p=0.68$. Overall, the students who received more exercises gamed significantly more often than the students who received fewer exercises, $\chi^2(1,N=388)= 24.33$, $p<0.001$.

4.2 Expressions of Anger

Overall, Scooter became angry considerably more often than he gave supplementary exercises. The median student saw an angry Scooter 13% of the time, and the student who saw an angry Scooter the most often saw an angry Scooter 38% of the time.

There did not appear to be an association between viewing an angry Scooter more often, and better learning. Students who received more expressions of anger did not have a significantly larger average learning gain than other students, whether we compared the top quartile to the other students, $t(37)=0.48$, $p=0.63$, effect size = 0.20σ, the top third, $t(37)=0.16$, $p=0.87$, or the top half, $t(37)=0.15$, $p=0.88$.

Additionally, there was no evidence of a relationship between Scooter's frequency of expressions of anger, and a reduction in gaming over time (as shown in Figure 5, right). Among the students who saw an angry Scooter the most often (top quartile), there was not a significant change either from day 1 to day 2, or day 2 to day 3, $\chi^2(1,N=79)= 0.04$, $p=0.84$, $\chi^2(1,N=50)= 0.83$, $p=0.36$.

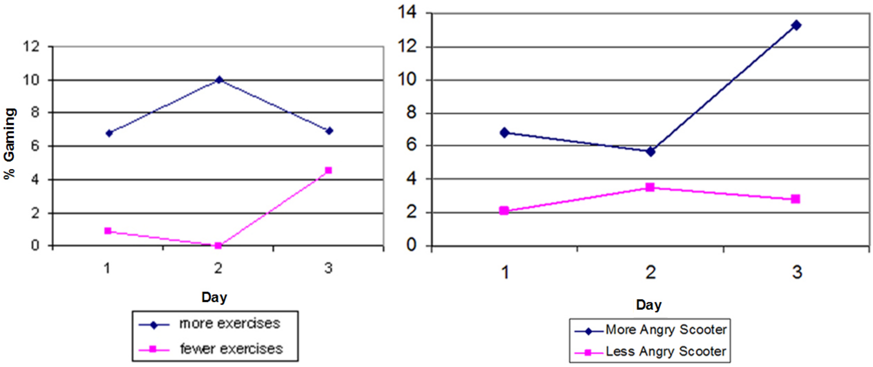


Fig. 5. Observed Gaming Over Time, in the Experimental Condition

5 Conclusions

In this paper, we present a re-designed tutor that responds to when students game the system, incorporating an animated agent, Scooter the Tutor. Students who received a large number of supplementary exercises from Scooter had high learning gains, and caught up to the rest of the class. This result is quite different from the pattern observed in the control condition and past studies [4,5], where students who game harmfully start out with lower pre-test scores, and fall further behind the rest of the class by the post-test.

Since students tend to game harmfully on the steps they know least well [4], the supplementary exercises may have been effective in large part because they offered additional learning support (and, perhaps, different learning support) for each student on the exact steps which that student found most difficult. Hence, we may be able to use a student's choice to game as an opportunity to learn more about where the student is having difficulty.

Incorporating Scooter into the tutor also led to about half as many students choosing to game. It is not entirely clear what aspect of the modified tutor led to the reduction in gaming. Neither students who saw an angry Scooter more often, nor students who received more supplementary exercises, reduced their gaming over time. One possibility is that simply knowing Scooter was present, and that he would make it impossible to hide gaming, led some students to game less. Thus, although Scooter's actions may not have directly affected the students who saw an angry Scooter, Scooter's presence may have motivated some students to avoid gaming during the entire lesson.

Overall, these results suggest that there is value to detecting and responding to differences in how students choose to use intelligent tutoring systems. By responding to gaming, we can develop tutors that help lower-performing students catch up to the rest of the class, and come closer to the goal of developing educational systems that help all students achieve.

Acknowledgements

We would like to thank Steve Ritter, Tom McGinnis, Sue Cameron, Russ Hall, Dina Crimone, Pat Battaglia, Shaaron Ainsworth, Cecily Heiner, Peter Scupelli, and Cristen Torrey for helpful suggestions and assistance. This work was funded by NSF grant REC-043779 to "IERI: Learning-Oriented Dialogs in Cognitive Tutors: Toward a Scalable Solution to Performance Orientation".

References

1. Aleven, V. (2001) Helping Students to Become Better Help Seekers: Towards Supporting Metacognition in a Cognitive Tutor. Paper presented at *German-USA Early Career Research Exchange Program: Research on Learning Technologies and Technology-Supported Education,* Tubingen, Germany.
2. Aleven, V., McLaren, B.M., Roll, I., Koedinger, K.R. (2004) Toward tutoring help seeking: Applying cognitive modeling to meta-cognitive skills. *Proceedings of the 7th International Conference on Intelligent Tutoring Systems (ITS 2004)*, 227-239.
3. Arroyo, I., Woolf, B. (2005) Inferring learning and attitudes from a Bayesian Network of log file data. *Proceedings of the 12th International Conference on Artificial Intelligence in Education,* 33-40.
4. Baker, R.S., Corbett, A.T., Koedinger, K.R. (2004) Detecting Student Misuse of Intelligent Tutoring Systems. *Proceedings of the 7th International Conference on Intelligent Tutoring Systems*, 531-540.
5. Baker, R.S., Roll, I., Corbett, A.T., Koedinger, K.R. (2005) Do Performance Goals Lead Students to Game the System? *Proceedings of the International Conference on Artificial Intelligence and Education (AIED2005)*, 57-64.

6. Baker, R.S., Corbett, A., Koedinger, K., Roll, I. (2005) *Detecting When Students Game The System, Across Tutor Subjects and Classroom Cohorts*. Proceedings of User Modeling 2005, 220-224.
7. Baker, R.S., Corbett, A.T., Koedinger, K.R., Wagner, A.Z. (2004) Off-Task Behavior in the Cognitive Tutor Classroom: When Students "Game The System". *Proceedings of ACM CHI 2004: Computer-Human Interaction*, 383-390.
8. Beck, J. (2005). Engagement tracing: using response times to model student disengagement. *Proceedings of the 12th International Conference on Artificial Intelligence in Education (AIED 2005)*, 88-95.
9. Cheng, R., Vassileva, J. (2005) Adaptive Reward Mechanism for Sustainable Online Learning Community. *Proc. of the International Conference on Artificial Intelligence in Education*, 152-159.
10. Klawe, M.M. (1998) Designing Game-based Interactive Multimedia Mathematics Learning Activities. *Proceedings of UCSMP International Conference on Mathematics Education.*
11. Microsoft Corporation. (1997) *Microsoft Office 97*. Seattle, WA: Microsoft Corporation.
12. Mostow, J., Aist, G., Beck, J., Chalasani, R., Cuneo, A., Jia, P., Kadaru, K. (2002) A La Recherche du Temps Perdu, or As Time Goes By: Where does the time go in a Reading Tutor that listens? Paper presented at *Sixth International Conference on Intelligent Tutoring Systems (ITS'2002).*
13. Murray, R.C., vanLehn, K. (2005) Effects of Dissuading Unnecessary Help Requests While Providing Proactive Help. *Proc. of the International Conference on Artificial Intelligence in Education*, 887-889.
14. Stevens, R., Soller, A., Cooper, M., & Sprang, M. (2004). Modeling the Development of Problem-Solving Skills in Chemistry with a Web-Based Tutor. *Proceedings of the 7th International Conference on Intelligent Tutoring Systems (ITS 2004)*, 580-591.