Squeezing out Gaming Behavior in a Dialog-Based ITS

Peter Hastings[1], Elizabeth Arnott[2], and David Allbritton[1]

[1] DePaul University
[2] Chicago State University

Abstract. Research Methods Tutor (RMT) is a dialog-based intelligent tutoring system which has been used by students in Research Methods in Psychology classes since 2003. Students interact with RMT to reinforce what they learn in class in five different topics. In this paper, we evaluate a different population of students and replicate our prior research: despite the relatively small amount of exposure during the term to RMT compared to other course-related activities, students learn significantly more on topics covered with RMT [1]. However, we did not find the same advantage for the dialog-based tutoring mode of RMT over the CAI mode. When transcript analyses indicated that a small but significant number of students were gaming the system by entering empty or nonsense responses, we modified the tutor to require reasonable attempts. This did lead some students to reform their gaming ways. In other cases, however, it resulted in disengagement from tutoring at least temporarily because reasonable answers were not recognized.

Key words: Dialog-based ITS, Gaming behaviors, Motivation

1 Introduction

Most collegiate Psychology programs require one or more classes in Research Methods [2]. Unfortunately, Psychology students find these classes to be especially difficult. The material is abstract and is normally learned by studying many specific cases of psychological research and then inferring general principles which will apply to their own experiments. We developed Research Methods Tutor (RMT), a dialog-based ITS, to help reinforce the concepts that students learn in research methods courses by engaging them in conversations about those topics. In previous research, we showed that RMT is effective. In this paper, we present data that replicates our previous results with a very different set of students. In related research, we identified a small, but significant number of students who were engaging in "gaming the system" behaviors. We also describe what happened when we modified RMT to encourage these students to re-engage with the tutor.

2 RMT in the field

2.1 The Tutor

RMT was modeled after the AutoTutor system [3] which was designed to follow the behavior of (non-expert) human tutors [4]. Some basic assumptions of this approach are:

- The tutor tends to control the dialog, using a variety of dialog moves to induce the student to provide particular information, and providing it when the student cannot.
- The tutor has a relatively shallow evaluation of the student's answers, and simply compares the student's response to the expected response to the current question.
- The tutor seldom gives direct negative feedback, instead preferring to simply give the expected answer or move to a related question.
- The tutor does not try to create an overall model of the student's knowledge, but comes to the tutoring session with a script of topics to cover.

These assumptions allow RMT to engage the student in extended conversations on the tutoring topics using relatively simple Natural Language Understanding techniques including LSA and keyword matching [5].

Research with AutoTutor has shown that it can produce remarkable learning effect sizes of up to 1σ [6]. However, most of this research has been done in a laboratory setting where research pool participants take a pretest, use the tutor intensively for some hours, and come back a week later for a second session of tutoring and the posttest. RMT was created first and foremost with the goal of providing additional support to our research methods students, and thus, our evaluations have differed significantly from the lab-based model.

Our participants take the pretest at the beginning of the term when they start their research methods class. In the first week of class, they are asked to login to RMT via the web to introduce themselves to the software, and install extensions for the agent-based version of the system if they can. (If they can't, they automatically use the text-only version of the system which provides the same information but doesn't use the talking head.) RMT includes five conceptual modules: Variables, Reliability, Validity, Experimental Design, and Ethics.[3] During the course of the term, the students are assigned to use RMT during the five weeks in which these topics are covered in class.

We have used two different types of control groups to assess the impacts of RMT. One control group is students in another section of Research Methods which is taught by the same instructor but without the use of RMT. Although students are (obviously) not randomly assigned to sections, we adopted this approach to minimize whatever carryover effects between conditions that might occur for students in the same sections. The other control condition (besides

[3] Additional modules are currently being developed, including statistics and graph interpretation.

no-tutoring) is a computer-aided instruction (CAI) condition in which students read (or are "read to" by the talking head) short passages of text and then take a few multiple choice questions. The text in the CAI condition was derived from the tutoring topics to ensure equivalent conceptual content. For each question answered in the CAI condition, the student is told whether the answer was right or wrong, but they are not told what their overall score was and are not required to achieve any particular level of performance in order to get credit for completing that CAI module. Students in the RMT classes are randomly split into two groups. One group gets tutoring for the first, third, and fifth modules, and CAI for the second and fourth, and the situation is reversed for the other group. At the end of the term, the students take a post-test which covers the original topics plus a transfer component.

2.2 Summary of prior results

In 2007, RMT was tested with almost 160 students from 5 sections of research methods classes at DePaul University. We compared the learning gains of tutored and non-tutored students, and performed a within-subjects comparison of tutoring versus CAI. The details were published in [7, 1] and are summarized here.

Using an ANCOVA with pretest score as covariate, and posttest score as the dependent variable, we found found that students in RMT classes scored significantly higher than students in control classes [$F(1, 155) = 23.21, p < .01$]. Using the learning effect size formula from the National Reading Panel [8], we calculated that the students in RMT classes learned 0.76σ more than control students. We were, frankly, astonished to see such a large effect size given the realities of our evaluation:

- The students only used the system for a combined total of 2–4 hours over the course of a ten-week term.
- They were interacting with the system primarily from their own homes or dorm rooms, often late at night, with (presumably) a range of distractions present.
- All the other class activities (lectures, tests, projects) may well have masked and/or interfered with whatever was learned from the tutor.

We concluded that RMT was very effective in *reinforcing* what the students were learning in the class, by having students engage in dialogs about those concepts.

In our within-subjects comparison of CAI versus tutoring, we found that students learned significantly more on topics on which they were tutored than on those on which they used CAI [$F(1, 71) = 4.627, p = .035$]. The NRP learning effect size was 0.34σ. We also checked if there were differences between students who used the agent-based mode of the system compared to the text-only mode. There was a marginally significant advantage of the agent-based mode [$F(1, 74) = 3.701, p = .058$]. This result must be interpreted with caution, however, since students essentially self-selected into this condition; if they couldn't follow the installation instructions or didn't have their own computer, they were put into the text-only condition.

2.3 New learning evaluations

In 2008, one of us took a faculty position at Chicago State University. This allowed us to attempt to replicate our evaluations of RMT with a very different group of students. Although both Chicago State and DePaul are located within Chicago, the student populations differ significantly. Table 1 summarizes a few differences.

Table 1. Comparison of student populations

DePaul University	Chicago State University
– 40% students of color – 77% of UG students are under 24 – 99% of incoming freshmen are under 21 – 75% of freshman live on campus	– 98% students of color – 60% have full-time jobs – 80% are parents – Average age of UG student = 26 – 95% of students live off-campus

Before evaluating the CSU students, we created a more concise version of the pre- and posttest. At 106 questions, our original test was rather onerous to the students, and we were concerned that they might not be trying their best on it — especially at the end of the term. The new version of the pretest had 50 questions, 10 per topic. The new posttest had the same questions plus five additional questions per topic as a transfer task. These questions presented experimental scenarios requiring more analytical than conceptual knowledge [9].

Students in the RMT condition ($n = 56$) took the tests and used the tutor or CAI as described above, with the exception that, as CSU runs on a semester schedule, the testing and tutoring took place over the course of 15 weeks instead of the 10 in DePaul's quarters. Students in the other section ($n = 31$) did not use RMT and served as the non-equivalent control group.

Again, we raised the same primary research question: Do students who use the tutor show higher learning gains from pretest to posttest than controls? We used an ANCOVA with the pretest score as the covariate, the condition (RMT, control) as independent variable, and posttest score as dependent variable. The results are shown in Table 2. The table gives the mean scores, standard deviations, and effect sizes for the RMT and control conditions on the first 50 questions of the posttest (identical to the pretest), the 25 transfer questions, and on the complete test.

Table 2. Evaluation results, 2008, CSU students

Test questions	Control	RMT	$F(1,84)$	Effect
First 50	19.8 (11.3)	32.5**(7.0)	54.78	1.39
Transfer (25)	9.1 (5.8)	12.9**(4.8)	14.05	0.71
Complete (75)	28.9 (16.7)	45.4**(10.5)	42.99	1.21

Significance level: $** \; p \leq .01$

Thus, for overall learning gains, we replicated our prior results showing that students learn significantly more when they use tutoring and CAI than when they do not, and achieved impressive effect sizes of 1.4σ, 0.7σ, 1.2σ on the basic test, transfer test, and complete test respectively.

We also addressed the question, does the dialog-based tutor result in higher learning gains on the posttest than the CAI version? Using a repeated measures ANOVA, we compared the scores for each student on tutor modules vs. CAI modules. Contrary to our prior results, there was no significant difference between the tutoring and CAI conditions $[F(1,27) = 3.202, p = 0.085]$. Tutor modules produced an average gain per module of 2.5. The average gain for CAI modules was 2.46.

Our third research question was: Does the agent result in greater learning gains on the posttest than text-only? Here, too, we found no significant differences between conditions with the CSU student population $[F(1,26) = 2.247, p = 0.146]$. There was a significantly smaller number of students using the agent condition at CSU (31% compared to 79% of the DePaul students). Two major factors could explain this: Microsoft has discontinued support for Microsoft Agents, and it doesn't work with the newer version of Internet Explorer. Fewer students had their own computers and were not allowed to install the software on lab computers.

Overall, we showed that use of RMT for tutoring and CAI does provide significant learning gains to students at Chicago State University. However, we did not find the advantage that we had found earlier for tutoring over CAI. One possible explanation that we wanted to explore was that these students were more adept in finding ways to "game the system". This topic is addressed in the next section.

3 Gaming behaviors

Identifying and counteracting gaming behaviors has become somewhat of a hot topic within the ITS community in recent years. When students "game the system," they typically focus their energies on finding ways to circumvent whatever pedagogical support the system was intended to provide. In this section, we describe some of the recent research in identifying and correcting gaming behaviors. Then we describe our analyses of gaming in RMT, and the steps that we took to counteract it.

3.1 Related work

Previous research in off-task or gaming behaviors in interactive learning environments has focused on five areas. Examples of the research findings follow:

1. Analyzing the effects of off-task or gaming behaviors on learning outcomes: Gaming was the only off-task behavior significantly correlated with learning gains [10]. Gaming has negative effects on learning both immediately, and in the aggregate [11].
2. Creating methods for automatically identifying off-task behaviors: Based on very fast actions, very slow actions, requests for help, and/or errors [12].
3. Determining features of individual learning problems that are correlated with gaming: [13] found only one of 79 features of cognitive tutor algebra problems that was significantly correlated with off task behaviors. Students went off-task much less when they were doing equation-solving. Other factors: abstract, ambiguous, or unclear problems [14].
4. Determining affective antecedents of gaming in participants: Students tend to game the system when they dislike the subject matter, have little educational self-drive, and are frustrated [15].
5. Trying to ascertain effective strategies for counteracting off-task behaviors. Better understanding of gaming *should* help reduce it [10–16], though there seems to be less empirical evidence supporting such claims.

While much of the recent research on off-task and gaming behaviors has been done within the context of cognitive tutors and the like, a notable exception is [16]. The authors call Crystal Island a Narrative-centered learning environment. It could also be called a serious game. Interestingly, gaming or off-task behavior within this type of game parallels that in the real world. Students may choose not to engage in goal-oriented behavior (according to the goal set in the game scenario), but instead to wander about, exploring the environment. This study used path analysis to differentiate goal-oriented and non-goal-oriented movements within Crystal Island.

3.2 Identifying gaming behaviors in RMT

In a dialog-based ITS, the student's actions are closer in some ways to those in a narrative-centered learning environment than in a traditional ITS. The student can enter absolutely any text in response to the tutor's questions or prompts. A cognitive tutor interface provides a limited number of actions. RMT's interface is exceedingly simple: besides the talking head or the text which present tutor utterances, there is only a text input box. What the student types into that box is only constrained by their educational motivation, their adoption of Gricean dialog maxims, and, of course, their understanding of the tutor's intentions and the intended answer.

While collecting materials for a large corpus analysis study, we noticed a small, but significant number of student transcript segments which indicated that the student was making a less-than-valiant attempt to answer the tutor. Some examples of such utterances are: “asdf,”, “j”, “hello” 60 times in a row, “help” 10 times in a row, “dude you voice is creepy,” “this is boring,” and “”.

RMT was designed to handle a range of different responses. In addition to student answers to its questions, RMT recognizes many different ways of asking it to repeat the question like (e.g. “what”, “come again”), statements about the student’s own comprehension (e.g. “dont know”, “do I need to know this?”), and questions about terminology. Capitalization and punctuation are ignored by the tutor and usually by the students. RMT includes an automatic spellchecker that attempts to map unrecognized words into those in its vocabulary. Because student spelling and word choice are so creative and because natural language processing in general is intractable, RMT attempts to “understand” student utterances that don’t fall into one of the categories above by comparing them to a small set of expected answers using LSA and keyword matching. This makes RMT fairly good at detecting *good* answers, but not so good at recognizing different types of (unexpected) bad answers.[4] In particular, RMT can not distinguish plain old bad answers from creative / unexpected good answers, other requests than those above, or random character strings. And because RMT is a helpful interlocutor, bad (non-good) answers prompt the tutor to provide the expected answer, and then ask a related question. Eventually, when the topic material has been covered by the tutor or the student, the tutor will provide a summary, and move on to the next problem.

We looked for evidence of gaming in transcripts of 234 students who used RMT between 2005 and 2009. Students were identified as extensive gamers if more than half of their utterances were blank, random strings, or non-responsive in some other way. Although our initial scan of transcripts indicated substantial gaming in about 10% of transcripts, only 15 out of 234 (6.4%) were labeled as extensive gamers. Seven more students showed significant but sub-threshold levels of gaming.

To examine the effects of gaming behavior on learning, we compared the learning outcomes of gamers and non-gamers. In marked contrast to previous research our data showed no significant effects of gaming on learning gains. One possible explanation is the great difference in the size of the two sets. Furthermore, students who gamed the tutor did not necessarily game the CAI modules. Module-by-module analyses showed no significant differences, but here the number of gamed tutor modules was even smaller than the number of gamers overall. If we combine the gamers with people who did not finish their modules we find that — although there aren’t significant differences, gamers and non-finishers together scored lower overall on all outcome measures. They scored significantly lower on the variables topic.

[4] RMT will trigger a remedial dialog for an expected bad answer.

3.3 Manipulating gaming

Although the overall extent of gaming was relatively small, it seemed both unnecessary and easily remedied (potentially). If RMT simply rejected answers with a similarity score of 0 to expected answers, then it could eliminate both empty and random responses and maintain its generous behavior for "nice tries". We added a third level (:ZERO) of evaluation for student answers, and altered the transition network which controls the tutors behavior. If a student answer got a :ZERO evaluation, the tutor would said something like, "I didn't get that" or "huh?", and repeat the previous question. We tested the system's behavior on a wide range of answers, and found that it was working as planned, so we included the modification in the online version of RMT approximately halfway through the Fall 2009 semester.

At the end of the semester, we examined the transcripts to see if we were successful in eliminating gaming. To our surprise (and embarrassment) the first thing we noticed was that a number of students required 2-3 times more turns to finish some of the topics. There were two culprits. One question had an obscure expected answer that LSA did not recognize as having any similarity with most student responses. Two questions expected numerical answers. Because LSA was trained to ignore numbers, our text pre-processor removed numbers along with punctuation before strings were spellchecked and compared to expected answers. When students reached this question, no answer they could give was accepted by RMT, and it continually repeated the same question. Thus, when we created one manipulation to attempt to reduce gaming behavior, we inadvertently created another that could frustrate students and increase gaming behavior.

3.4 Results

This section describes our analyses of the results of this dual manipulation of gaming behavior. As with the overall comparison of gamers and non-gamers across the different terms, the students in Fall 2009 showed no significant effects of gaming on learning gains. Again a relatively small number of students (4 of 39) provided a significant number of non-responsive answers. For two of these students, the modification of the tutor's behavior appeared effective in eliminating gaming behavior. On the topics completed before the modification, both students entered primarily blank or random answers. After the modification, they answered the questions.

For the problematic questions where the tutor accepted few or no answers, we coded the students as "frustrated" if after a number of attempts, they began to enter blank or random responses. Although 6 students appeared to engage in gaming behavior when frustrated in this way, 4 of them subsequently completed other tutor modules without gaming. Furthermore, students who were "frustrated" did not score differently on any of the subtopics or the posttest overall.

Although there were no significant differences in learning gains between CAI and tutoring conditions for the CSU students as a whole, in Fall 2009, with the anti-gaming manipulation, students learned significantly more from the CAI modules than they did from the tutoring modules. A repeated measures ANOVA comparing each participant's scores for the tutor modules to that same participant's score for the CAI modules showed that the scores on the CAI modules were significantly higher, $[F(1,21) = 7.299, p = 0.013]$, tutoring mean gain = 2.4 ($SE = .34$), CAI mean gain = 4.2 ($SE = .91$).

Although fewer students completed the tutoring topics covered later in the Fall 2009 semester, this pattern was noted in other terms as well. There was no significant difference in the rates of topic completion between the terms.

4 Conclusions and future work

In this paper, we described learning results for RMT in two different student populations. Significantly, these learning gains were recorded in everyday use of the system, not in a laboratory context. We also described our analyses of gaming behavior in RMT and our (somewhat unfortunate) attempts to deter gaming. Analyses of the student transcripts showed that the change did, in fact, lead to reduction in gaming behavior in some, but not all students who had previously started gaming the system.

We also had the opportunity to analyze the effects of increased frustration on users of a dialog-based tutor. Although some students did disengage, for most it was only temporary. When they went on to other topics, they went back to interacting with the tutor as they had before.

Although it was too late to help the Fall 2009 students, the "number problem" was easily fixed, and RMT now accepts numerical answers. The tutor's behavior now makes it easier to identify obscure expected answers as well. Previously, the tutor's behavior wasn't markedly different for problematic errors. Now, we can integrate triggers into RMT that identify when students get stuck on a particular question, allowing the student to continue on, and alerting us that the expected answers may need to be changed.

In future work, we would also like to explore the possibility of giving the student some indicator of the cummulative quality of their responses. We hope that this could make it more clear to the students the relationship between the effort they put into answering the questions, and the efficiency with which they move through the tutoring topics. We would also like to develop a test harness for the system. This is will be a challenge, however, due to the natural language input to the system, and the dynamic determination of response and dialog direction.

References

1. Arnott, E., Hastings, P., Allbritton, D.: Research Methods Tutor: Evaluation of a dialogue-based tutoring system in the classroom. Behavior Research Methods **40**(3) (2008) 694–672

2. Perlman, B., McCann, L.I.: The structure of the psychology undergraduate curriculum. Teaching of Psychology **26** (1999) 171–176
3. Graesser, A., Person, N., Harter, D., the TRG: Teaching tactics and dialog in AutoTutor. International Journal of Artificial Intelligence in Education **12** (2001) 23–39
4. Person, N.K.: An analysis of the examples that tutors generate during naturalistic one-to-one tutoring sessions. PhD thesis, University of Memphis, Memphis, TN (1994)
5. Wiemer-Hastings, P., Allbritton, D., Efron, J., Arnott, E.: Research methods tutoring in the classroom. In: AIED2003 - Supplementary Proceedings of the 11th International Conference on Artificial Intelligence in Education, Sydney, University of Sydney (2003) 388 – 392
6. Graesser, A., Jackson, G., Mathews, E., Mitchell, H., Olney, A., Ventura, M., Chipman, P., Franceschetti, D., Hu, X., Louwerse, M., Person, N., TRG: Why/AutoTutor: A test of learning gains from a physics tutor with natural language dialog. In: Proceedings of the 25th Annual Conference of the Cognitive Science Society, Mahwah, NJ, Erlbaum (2003)
7. Arnott, E., Hastings, P., Allbritton, D.: RMT in the classroom. In: Proceedings of the Midwest Artificial Intelligence and Cognitive Science conference. (2007)
8. National Reading Panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implicaitons for reading instruction. Technical Report NIH 00-4754, National Institute of Child Health & Human Development, Washington, DC (2008)
9. Bloom, B., ed.: Taxonomy of educational objectives: The classification of educational goals: Handbook I, cognitive domain. Longmans, Green, New York (1956)
10. Baker, R.S., Corbett, A.T., Koedinger, K.R., Wagner, A.Z.: Off-task behavior in the cognitive tutor classroom: When students “game the system”. ACM CHI 2004: Computer-Human Interaction (2004) 383–390
11. Cocea, M., Hershkovitz, A., Baker, R.: The impact of off-task and gaming behaviors on learning: Immediate or aggregate? In Dimitrova, V., Mizoguchi, R., du Boulay, B., eds.: Proceedings of the 14th International Conference on Artificial Intelligence in Education, Amsterdam, IOS Press (2009) 507–514
12. Baker, R.S.: Modeling and understanding students’ off-task behavior in intelligent tutoring systems. In: CHI ’07: Proceedings of the SIGCHI conference on Human factors in computing systems, New York, NY, USA, ACM (2007) 1059–1068
13. Baker, R.: Differences between intelligent tutor lessons, and the choice to go off-task. In: Proceedings of the 2nd International Conference on Educational Data Mining. (2009) 11–20
14. Baker, R.S., de Carvalho, A.M., Raspat, J., Aleven, V., Corbett, A.T., Koedinger, K.R.: Educational software features that encourage and discourage “gaming the system”. In Dimitrova, V., Mizoguchi, R., du Boulay, B., eds.: Proceedings of the 14th International Conference on Artificial Intelligence in Education, Amsterdam, IOS Press (2009) 507–514
15. Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., Koedinger, K.: Why students engage in “gaming the system” behavior in interactive learning environments. Journal of Interactive Learning Research **19**(2) (2008) 185–224
16. Rowe, J.P., McQuiggan, S.W., Robison, J.L.: Off-task behavior in narrative-centered learning environments. In Dimitrova, V., Mizoguchi, R., du Boulay, B., eds.: Proceedings of the 14th International Conference on Artificial Intelligence in Education, Amsterdam, IOS Press (2009) 99–106