Data-driven causal modeling of “gaming the system” and off-task behavior in Cognitive Tutor Algebra

Stephen E. Fancsali
Carnegie Learning, Inc.
437 Grant Street, Suite 918
Pittsburgh, PA 15219, USA
sfancsali@carnegielearning.com

Abstract

“Gaming the system” and off-task behavior in intelligent tutoring systems (ITSs) have been found to be negatively associated with student learning outcomes. We summarize recent work to determine whether these behaviors are likely *causes* of decreased learning in the Cognitive Tutor® Algebra ITS using algorithmic search for the structure of graphical causal models. We apply data-driven, software “detectors” of these behaviors to observed log data for 102 adult learners in an algebra course at the University of Phoenix®. We find evidence that “gaming the system” is a cause of decreased learning and that, while the two are correlated, off-task behavior is not a cause of decreased learning. A posited measure of the extent to which students manifest “known misconceptions” in the tutor mediates the causal link between gaming the system and learning. We discuss these results and several future research topics.

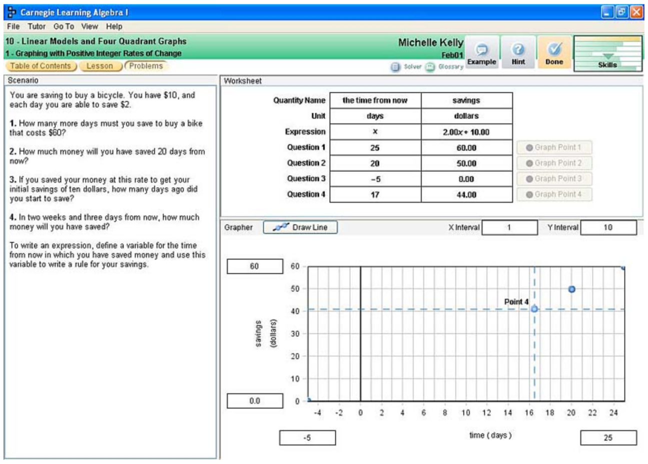
1 Introduction

Educational data mining and learning science researchers have developed a variety of sensor-free, data-driven, software “detectors” of student behavior and affect while interacting with intelligent tutoring systems (ITSs) (e.g., [1]-[7]). Many such detectors focus on inferring, from ITS log data, whether students are likely to be engaging in “gaming the system” (e.g., [1] [3]) and/or off-task behavior (e.g., [2]). While these behaviors are found to be correlated with decreased learning in ITSs (e.g., [8]), we summarize recent work [9]-[10] that uses the framework of data-driven search for graphical causal models [11]-[12] to provide evidence that the correlation between “gaming the system” behavior and decreased learning, observed in Carnegie Learning’s Cognitive Tutor® (CT) software [13], is attributable to a causal relationship. Further, we find no evidence that the correlation of off-task behavior and decreased learning arises from a causal relationship in our sample of 102 adult learners in an algebra course at the University of Phoenix® (UoP). We show that the extent to which students produce errors that reflect possible “known-misconceptions” (i.e., that trigger immediate tutor feedback messages) in the CT is one mediator of the link between gaming the system behavior and decreased learning and conclude by discussing our results and future research.

1.1 Cognitive Tutor

Carnegie Learning’s CT is an ITS used by hundreds of thousands of mathematics learners, ranging from middle schoolers to college-level undergraduates, every year, in the United States and abroad. CT divides mathematics curricula into topical units, comprised of sections covering relatively fine-grained sub-topics. In each section, students are presented a number of multi-part problems; each part of a problem presents the student an opportunity to practice particular knowledge components (KCs) or skills, into which the mathematics domain has been atomized. The learner’s action to complete part of a problem (e.g., filling in a text field in the table for the problem in Figure 1) is assessed by the CT as either correct or incorrect. The CT responds immediately with its assessment of the learner's action; sometimes the CT recognizes incorrect student actions (i.e., inputs) as "known misconceptions" (e.g., transforming a negative number into a positive number). In addition to responding with whether an action is correct or incorrect, the CT responds to actions that it tracks as misconceptions with "just-in-time" (JIT) feedback messages. Learners may request help from the CT while solving any part of a problem.

CT adapts to students' knowledge by tracking progress as students are presented opportunities to practice particular KCs using a probabilistic framework called Bayesian Knowledge Tracing (BKT) [14]. The CT judges that a student has mastered a KC when its estimate of the probability of a student's knowledge of a KC exceeds 95%. CT adaptively presents problems to students that provide opportunities to practice KCs within a particular section that a student has yet to master. When a student is judged to have mastered all KCs associated with a particular section, the student "graduates" to the next section (or unit, if graduating from a unit's last section).



10 - Linear Models and Four Quadrant Graphs
1 - Graphing with Positive Integer Rates of Change

Michelle Kelly
Feb01

You are saving to buy a bicycle. You have $10, and each day you are able to save $2.

1. How many more days must you save to buy a bike that costs $60?

2. How much money will you have saved 20 days from now?

3. If you saved your money at this rate to get your initial savings of ten dollars, how many days ago did you start to save?

4. In two weeks and three days from now, how much money will you have saved?

To write an expression, define a variable for the time from now in which you have saved money and use this variable to write a rule for your savings.

Quantity Name	the time from now	savings
Unit	days	dollars
Expression	x	2.00x + 10.00
Question 1	25	60.00
Question 2	20	50.00
Question 3	-5	0.00
Question 4	17	44.00

X Interval 1
Y Interval 10

60
0.0
-5
25

Figure 1: Screenshot of Cognitive Tutor Algebra.

1.2 Detectors of "gaming the system" and off-task behavior in an ITS

"Gaming the system" and off-task behavior in ITSs like the CT are the subject of a great deal of recent research. Gaming the system refers to learner behavior that involves taking advantage of ITS properties to progress through material without genuine learning [15]. Generally, two rough types of behavior comprise gaming the system [16]: " hint-abuse," [17] whereby a student rapidly uses an ITS's hint facility, sometimes reaching a "bottom-out" hint that provides a correct answer, and rapid and/or systematic guessing until the student provides a correct answer. Off-task behavior refers to learner disengagement from the ITS environment and learner actions that are not related to learning tasks [2]. Both types of behavior have been found to be negatively associated with aggregate learning outcomes (e.g., [8], [16]).

Given this negative association, descriptions of "harmful" gaming the system behavior in the literature generally use that moniker in a non-causal way. For example, Baker, et al. [16] note: "The word 'harmful' is used for brevity and simplicity in discussion; there is still not conclusive evidence as to whether the relationship between harmful gaming and learning is causal or correlational" [16, pg. 291, footnote]. Further, some gaming the system behavior is described as "non-harmful" and is not associated with decreased learning [1], [16]. For example, behavior that on the surface appears to be gaming the system may in fact correspond to students seeking bottom-out hints as worked examples [18].

Sensor-free, software-based detectors of gaming the system [3], off-task behavior [2], and affect [7] have been developed that use fine-grained logs of student actions to yield predictions of whether (sets of) student interactions correspond to such behavior or to a learner being in one of several affective states. Such predictions have been validated by their correspondence to field observations[1] of student interaction with ITSs, like the CT, in a computer lab or classroom. We focus on gaming the system and off-task behavior, detectors of which use Latent Response Models [20] specified using various features "distilled" from logs from CT (or a similar ITS) [16]. The detectors generate a numerical value for each learner action, to which a threshold is applied to determine whether any particular action is likely an example of gaming the system or off-task behavior. Following work that demonstrates negative correlations of both gaming the system and off-task behavior with aggregate learning outcomes, we investigate whether gaming the system and off-task behavior are causally related to decreased learning.

2 Graphical causal models

To answer such a causal question from non-experimental data, we turn to the framework of graphical causal models [12], and specifically, structure-learning algorithms[2] to infer (equivalence classes of) causal models from observational data under various conditions [11]. Of much interest to statisicians, computer scientists, and philosophers over the past 20+ years, such models frequently take the form of directed acyclic graph (DAG) causal models (e.g., Bayesian networks, when coupled with an associated probability distribution) and have been successfully used to model a variety of educational data (e.g., [21]-[23]). In such models, nodes represent observed variables, and edges connecting nodes represent causal relationships between variables.

To reason about systems in which there may be unmeasured (i.e., confounding) common causes of measured variables (i.e., the situation we face in most real-world, scientific applications), we can use graphical objects called Partial Ancestral Graphs (PAGs) [24], [11]. The FCI algorithm [11] infers a PAG causal model from conditional independence[3] relationships among observed or measured variables while allowing unmeasured common causes of measured variables. Assuming there is no selection bias, there are four types of edges in a PAG between any two distinct variables X and Y, that we interpret in the following ways:

- X o→ Y : Y is not an ancestor of X in any graph in the equivalence class represented by the PAG. Thus, in every graph in the equivalence class: X is an ancestor (i.e., cause) of Y, or X and Y share an unmeasured common cause (or both).
- X o−o Y : (1) X is an ancestor of Y, (2) Y is an ancestor X, (3) X and Y have an unmeasured common cause, or (4) either (1) & (3) or (2) & (3).
- X ↔ Y : X and Y share an unmeasured common cause in every member of the equivalence class.
- X → Y : X is an ancestor of Y. That is, X is a cause of Y, but X need not be a direct cause of Y.

We apply the FCI algorithm to data that we now describe.

3 Data

We analyze data from 102 non-traditional, adult learners at UoP that used CT Algebra in 2010 in an introductory, undergraduate algebra course for students seeking four year degrees. Students were drawn from both online and on-campus offerings of this course at UoP. We consider log data from the last of four modules of units in this course. This module includes units on the following five algebra topics:

- Systems of Linear Equations
- Systems of Linear Equations Modeling

[1] using, for example, the Baker-Rodrigo Observation Method Protocol (BROMP) [19]

[2] most of which are freely available in the TETRAD suite of software: http://www.phil.cmu.edu/projects/tetrad/

[3] See [11] for details on the two underlying assumptions that connect conditional independence and causal structure: the Causal Markov Condition and the Causal Faithfulness Condition.

- Linear Inequalities
- Graphing Linear Inequalities
- Systems of Linear Inequalities

CT log files represent "tutor transactions," a combination of the student's action on a part of a problem and the tutor's response to that action. To these data, we apply aforementioned detectors of gaming the system [3] and off-task behavior [2] and construct variables for modeling that we describe in the following section. Data also include a pre-test score for this module of units. The learning outcome we consider is each student's final exam score for the entire course.

4 Analysis & results

While many questions remain about how to best construct variables or extract features from fine-grained log files to represent aggregate learner behaviors, affective states, and other important (high-level or aggregate) education features (cf. [10]), we construct variables over the module of CT Algebra following those provided in previous analysis of the negative association of aggregate gaming the system and off-task behavior with learning outcomes [8]. These variables are defined in terms of tutor "steps," here defined as consecutive sequences of user actions practicing the same KC. To these variables, we add a count of the number of user actions that triggered JIT feedback plus the aforementioned pre-test score and course final exam score. Adding the count of learner actions that produce JIT feedback to the model is suggested both by the conceptualization of gaming the system and by a data-driven procedure detailed in [10]. The variable names and explanations are as follows:

- PRETEST_SCORE: module pre-test score
- STEPS_OFFTASK: total number of steps in module that contained at least one action judged likely to be off-task
- STEPS_GAMED: total number of steps in module that contained at least one action judged likely to be an example of gaming the system
- TOTAL_STEPS: number of steps required for student to work through course material
- ACTIONS_PRODUCING_JIT_FEEDBACK: number of actions in module that produce JIT feedback (i.e., known misconceptions)
- FINAL_EXAM: score on UoP algebra course final exam

FCI can account for background knowledge, including the temporal order of variables. For example, the pre-test precedes use of the tutor, so STEPS_GAMED cannot be a cause of PRETEST_SCORE. If PRETEST_SCORE and STEPS_GAMED are correlated, then PRETEST_SCORE causes STEPS_GAMED, or they share at least one common cause (or both). In our model, we impose a time-ordering on these variables: PRETEST_SCORE is temporally prior to the four behavior variables that track gaming the system, off-task behavior, total steps, and learner actions that produce or trigger JIT feedback, and all of these variables are temporally prior to FINAL_EXAM. The result of FCI search over these variables is the PAG of Figure 2. If we generate a DAG from the PAG of Figure 2 by orienting each o→ edge as a directed → edge, we can estimate a structural equation model[4]; the signs of path coefficients of this estimated model are provided in Figure 2.

Notably, STEPS_GAMED is inferred to be an ancestor (i.e., cause) of ACTIONS_PRODUCING_JIT_FEEDBACK which is an inferred cause of FINAL_EXAM. Since STEPS_GAMED and ACTIONS_PRODUCING_JIT_FEEDBACK are positively correlated, and the latter variable and our learning outcome, FINAL_EXAM, are negatively correlated, we find evidence that the negative correlation of gaming the system behavior and our learning outcome is due to a causal relationship, even allowing that there may be unmeasured common causes of measured variables in our analysis. That ACTIONS_PRODUCING_JIT_FEEDBACK for a student mediates the causal relationship between gaming and learning may prove useful, as it is a measure that is easily gleaned from CT log data. Further, we find that STEPS_OFFTASK is not a cause of any of the other variables in our model. Thus, the (negative) correlation of off-task behavior and our

[4] This model fits the data according to a statistical test comparing the implied covariance matrix of the model with the observed covariance matrix ($\chi^2(8) = 13.78$, p = .09) [25].

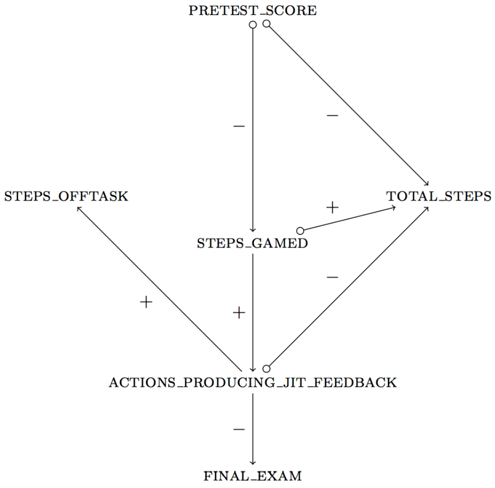


Figure 2: PAG model for UoP CT Algebra data

learning outcome does not arise out of a causal relationship. Plausibly, students with higher pre-test scores for this module tend to require fewer overall steps in solving problems to graduate from the module's sections and units, and they tend to engage in less gaming the system behavior. From data alone, we are unable to infer whether these relationships are causal.

5 Discussion & future work

We provide evidence for a causal link between gaming the system in an ITS and decreased learning. If teachers focus on reducing gaming the system, this is likely to produce better learning outcomes than focusing on reducing student off-task behavior; given our results, intervening on off-task behavior does not get at a root cause of poor performance. We have generally conceptualized learner actions tracked by CT that produce JIT feedback, the mediator in the causal chain linking gaming the system and decreased learning, as reflecting misconceptions, but they may simply reflect "shallow" answers (e.g., any number that appears in a problem) that come from systematic/rapid guessing associated with gaming the system. Further investigation should shed light on this mechanism.

Many questions remain for applying methods of causal structure search to educational data sets. Given the advent of "big data" in education (e.g., MOOCs and widely deployed courseware like the CT), how do we best construct aggregate variables to represent student-level features that are meaningful and appropriate for causal modeling (e.g., that preserve conditional independence among phenomena we model) (cf. [10])? Further, with the development of detectors of student affect in ITSs [7], we enter the realm of modeling latent phenomena; specifying appropriate measurement models, constructing proxies for (or explicitly modeling) latent constructs (e.g., boredom and frustration), and incorporating these variables into causal model search (i.e., providing integrated causal models of student behavior, affect, and learning) remains a topic for future research.

Acknowledgments

The author gratefully acknowledges Ryan S.J.d. Baker, Susan Berman, David Danks, Sujith Gowda, Tristan Nixon, Steven Ritter, and Richard Scheines for their comments and assistance. As noted, portions of these results were previously reported [9]-[10] and completed while the author was at Carnegie Mellon University's Department of Philosophy.

References

[1] 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.

[2] Baker, R.S.J.d. (2007). Modeling and Understanding Students' Off-Task Behavior in Intelligent Tutoring Systems. *Proceedings of the 2007 Conference on Human Factors in Computing Systems*, 1059-1068.

[3] Baker, R.S.J.d., de Carvalho, A. M. J. A. (2008). Labeling Student Behavior Faster and More Precisely with Text Replays. *Proceedings of the 1st International Conference on Educational Data Mining*, 38-47.

[4] Beck, J. (2005). Engagement Tracing: Using Response Times to Model Student Disengagement. *Proceedings of the 12th International Conference on Artificial Intelligence in Education*, 88-95.

[5] Walonoski, J.A., Heffernan, N.T. (2006). Detection and Analysis of Off-Task Behavior in Intelligent Tutoring Systems. *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, 382-391.

[6] Beal, C.R., Qu, L., Lee, H. (2006). Classifying Learner Engagement Through Integration of Multiple Data Sources. *Proceedings of the 21st AAAI Conference on Artificial Intelligence*, 151-156.

[7] Baker, R.S.J.d., Gowda, S.M., Wixon, M., Kalka, J., Wagner, A.Z., Salvi, A., Aleven, V., Kusbit, G., Ocumpaugh, J., Rossi, L. (2012). Sensor-free Automated Detection of Affect in a Cognitive Tutor for Algebra. *Proceedings of the 5th International Conference on Educational Data Mining*, 126-133.

[8] Cocea, M., Hershkovitz, A., Baker, R.S.J.d. (2009). The Impact of Off-Task and Gaming Behavior on Learning: Immediate or Aggregate? *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, 507-514.

[9] Fancsali, S.E. (2012). Variable Construction and Causal Discovery for Cognitive Tutor Log Data: Initial Results. *Proceedings of the Fifth International Conference on Educational Data Mining*, 238-239.

[10] Fancsali, S.E. (2013). *Constructing Variables that Support Causal Inference*. Ph.D. Thesis, Department of Philosophy, Carnegie Mellon University.

[11] Spirtes, P., Glymour, C., Scheines, R. (2000). *Causation, Prediction, and Search*. 2nd Edition. Cambridge, MA: MIT Press.

[12] Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. 2nd Edition. Cambridge: Cambridge UP.

[13] Ritter, S., Anderson, J.R., Koedinger, K.R., Corbett, A.T. (2007). Cognitive Tutor: Applied Research in Mathematics Education. *Psychonomic Bulletin and Review*, 14(2), 249-255.

[14] Corbett, A.T., Anderson, J.R. (1995). Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User Adapted Interaction*, 4, 253-278.

[15] Baker, R.S.J.d., Corbett, A.T., Koedinger, K.R., Evenson, S.E., Roll, I., Wagner, A.Z., Naim, M., Raspat, J., Baker, D.J., Beck, J. (2006). Adapting to When Students Game an Intelligent Tutoring System. *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, 392-401.

[16] Baker, R.S.J.d., Corbett, A.T., Roll, I., Koedinger, K.R. (2008). Developing a Generalizable Detector of When Students Game the System. *User Modeling and User-Adapted Interaction*, 18, 287-314.

[17] Aleven, V., Koedinger, K. R. (2000). Limitations of Student Control: Do Students Know When They Need Help? *Proceedings of the 5th International Conference on Intelligent Tutoring Systems*, 292-303.

[18] Shih, B., Koedinger, K.R., Scheines, R. (2011). A Response-Time Model for Bottom-Out Hints as Worked Examples. In: C. Romero, S. Ventura, M. Pechenizkiy, and R.S.J.d. Baker (eds.) *Handbook of Educational Data Mining*, 201-211. Boca Raton, FL: CRC Press.

[19] Ocumpaugh, J., Baker, R.S.J.d., Rodrigo, M.M.T. (2012). Baker-Rodrigo Observation Method Protocol (BROMP) 1.0. Training Manual version 1.0. Technical Report. New York, NY: EdLab. Manila, Philippines: Ateneo Laboratory for the Learning Sciences.

[20] Maris, E. (1995). Psychometric Latent Response Models. *Psychometrika*, 60, 523-547.

[21] Scheines, R., Leinhardt G., Smith, J., Cho, K. (2005). Replacing Lecture with Web-Based Course Materials. *Journal of Educational Computing Research*, 32, 1-26.

[22] Rai, D., Beck, J.E. (2011). Exploring User Data from a Game-Like Math Tutor: A Case Study in Causal Modeling. *Proceedings of the 4th International Conference on Educational Data Mining*, 307-311.

[23] Rau, M., Scheines, R., Aleven, V., Rummel, N. (2013). Does Representational Understanding Enhance Fluency - Or Vice Versa? Searching for Mediation Models. *Proceedings of the 6th International Conference on Educational Data Mining*, 161-168.

[24] Richardson, T.S. (1996). A Discovery Algorithm for Directed Cyclic Graphs. *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*, 454-461.

[25] Bollen, K. (1989). *Structural Equations with Latent Variables*. Wiley.