Confounding Carelessness? Exploring Causal Relationships Between Carelessness, Affect, Behavior, and Learning in Cognitive Tutor Algebra

Stephen E. Fancsali
Carnegie Learning, Inc.
437 Grant Street, 20th Floor
Pittsburgh, PA 15219 USA
1.888.751.8094 x219
sfancsali@carnegielearning.com

ABSTRACT

Studies have found positive correlations between affective states (e.g., confusion, boredom) and learning outcomes in educational technologies like ASSISTments and Carnegie Learning's Cognitive Tutor. The adage that "correlation does not imply causation" is especially apt in light of these observations; it seems counterintuitive that increasing student boredom or confusion (e.g., designing systems that bore or confuse students) will benefit learning. One hypothesis to explain positive correlations between boredom and learning suggests that carelessness is a "confounding" common cause of boredom and another construct linked to learning. We consider a Cognitive Tutor Algebra dataset in which boredom and confusion are positively correlated with learning. Prior causal modeling of this data suggests that various behavioral and affective features (e.g., boredom and gaming the system) share unmeasured common causes. We provide a correlational analysis and causal models of this data that situate carelessness among behaviors and affective states to determine whether (and how) carelessness plays a confounding role.

Keywords

causal models, causal discovery, structural equation modeling, carelessness, boredom, confusion, affect, gaming the system, off-task behavior, Cognitive Tutor, intelligent tutoring systems

1. INTRODUCTION

Recent research in educational data mining has led to the development of sensor-free, data-driven approaches to "detect" various behavioral and affective features from logs of learner interactions with technologies like intelligent tutoring systems (ITSs). Since such approaches to detecting phenomena like "gaming the system" [3-4], off-task behavior [1], and affective states [5] have been validated against field observations of learner behavior, a natural next step for researchers has been to use the predictions of detectors as inputs to predictive models of substantive learning outcomes in what have been called "discovery with models" approaches [6]. Such approaches have sought to answer questions about whether the tendency of learners to game the system or become bored using a system are predictive of outcomes like post-tests and standardized test scores (e.g., [11, 17]).

Our recent work [13] advocates seeking causal knowledge about learner behavior, affect, and learning, even when faced with non-experimental data, and that graphical causal models and data-driven search for their structure [22] provide an avenue for *causal* discovery with models. Findings using data from Carnegie Learning's Cognitive Tutor (CT) ITS [19] suggested that most affect and behavior variables shared unmeasured common causes.

The present work integrates detectors of carelessness into this work [13]. Carelessness was correlated with a variety of affective phenomena in an ITS with features similar to CT (e.g., [21]) and has been hypothesized to play a causal role among affective states as well (e.g., as a cause of boredom [17] or effect of boredom [8]). Other work emphasizes relationships between engagement and carelessness [9-10]. Our findings suggest a causal link between concentration and carelessness and possible causal links between confusion, gaming the system, and carelessness.

2. PRELIMINARIES

2.1 Motivation & Outline

Recent studies (e.g., [13, 17]) observe positive correlations between learning and the propensity to be in affective states like boredom and confusion, but it seems counterintuitive that increasing student boredom or confusion is likely to benefit learning (i.e., that these correlations are because of causal links). Several hypotheses have been proffered to explain such positive correlations. One hypothesis, for the ASSISTments system [14], is that learners become bored when they make careless mistakes and are required to work through step-by-step breakdowns of math problems [17]; learners with greater knowledge are more likely to be careless and bored, but since they are capable learners they will have better learning outcomes, providing a possible explanation for a positive correlation between boredom and learning.

Further, causal models of affect, behavior and learning in CT Algebra finds that boredom and gaming the system behavior are negatively correlated and suggest that they share an unmeasured (or latent) common cause (i.e., a "confounding" variable) [13]. Boredom's negative correlation with gaming the system, and gaming behavior's negative correlation with learning helped to explain the overall positive correlation of boredom and learning in that study. This same study also found a positive correlation between confusion and learning, and causal models suggested that confusion and gaming may be confounded. The hypothesis of [17] about carelessness may be appropriate for CT; incorrect responses despite knowledge will lead students to be presented more practice on skills they already know because CT will decrease its

estimate of skill mastery based on incorrect responses, and this could lead to boredom.

Recent work has focused on modeling carelessness [20] in systems like ITSs by using context-sensitive models to predict whether particular incorrect responses are likely examples of "slips" in which students answer incorrectly despite knowing a skill [2], and have explored correlations between carelessness and affective states (e.g., [21]).

We now introduce CT Algebra and detector models. We then review graphical causal models and data-driven structure search before explaining prior work and presenting novel causal models that incorporate carelessness. We conclude with discussion.

2.2 Cognitive Tutor (CT) Algebra

Carnegie Learning's CT is an ITS for mathematics used by hundreds of thousands of learners every year across the United States and internationally. CT breaks down mathematics subject areas like algebra into fine-grained skills or knowledge components (KCs), the mastery of which is used to determine learner progress through a series of topical sections that comprise broader units. Each section is comprised of multi-step problems that allow for the assessment of student progress toward mastery of fine-grained KCs.

CT assesses KC mastery using a probabilistic framework called Bayesian Knowledge Tracing (BKT) [12]. BKT assesses learner progress to mastery by assuming that a learner is either in the "unknown" state for a KC or the "known" state for a KC (i.e., KC mastery) and uses observations of practice opportunities for each KC to predict the state of a learner is at any given time. To make this prediction, BKT provides for four parameters for each KC: (1) the probability of prior knowledge or mastery of the KC, (2) the probability of a transition from the unknown to the known state at a given KC practice opportunity, (3) the probability that a learner guesses (i.e., is in the unknown state but answers correctly), and (4) the probability that the learner "slips" (i.e., has mastered a KC but provides an incorrect response).

2.3 Affect, Behavior, & Carelessness

Educational data mining researchers seek to avoid obtrusive, costly, and non-scalable sensor-based methods for measuring learner (dis-) engagement and affect with systems like ITSs by developing data-driven predictive models, frequently referred to as "detectors" that rely only on features that can be "distilled" from fine-grained log data. Detector models use machine learning methods applied to distilled features to make predictions about whether particular learner interactions with a system are likely to be instances of particular types of behavior. Detector models are validated against field observations in real classrooms. For correlational and causal modeling, we quantify levels of behavior per student by calculating the proportions of problem-solving steps deemed to be likely the result of behaviors like gaming the system or off-task behavior, which we now briefly explicate.

Gaming the system [3-4] refers to behavior in which learners attempt to make progress through content without genuinely learning or mastering appropriate skills (e.g., by incorrectly providing numbers within problem statements). A robust finding of previous efforts is that there is evidence that gaming the system is a cause of decreased learning [13]. Off-task behavior refers to learner disengagement from the learning environment and learning [1]. Recent efforts did not find evidence for a causal link between off-task behavior and learning.

Evidence also suggests that affective states play an important role in learning (e.g., [18]). Detector models similar to those for gaming the system and off-task behavior have been developed for affective states like boredom, confusion, and engaged concentration [5]. Modeling efforts for a CT Algebra dataset provided a somewhat complicated causal picture; while boredom and confusion may be *negatively* correlated with another factor that causes *decreased* learning, gaming the system, (hence positively correlated with learning), there are likely unmeasured common causes of these states and gaming the system.

Learner carelessness has been discussed as problematic in classrooms since at least the 1950s [21]. Other work identifies carelessness as a problem even among high-performing students [9-10]. Recent work on data-driven detector models [20] seeks to operationalize carelessness by focusing on the notion of "slipping," when learners answer incorrectly despite knowing a skill. In standard BKT, the parameter for slipping remains constant per KC over time; contextual models of guessing and slipping predict whether particular correct and incorrect responses are likely the result of learners guessing or slipping based on aspects of their performance [2]. The contextual slip model that predicts whether particular incorrect responses are instances of slipping is built in the same manner as other detector models. Operationalized as contextual slipping, carelessness can be quantified on a per learner basis by calculating the mean probability with which contextual slip models predict that incorrect actions are examples of slipping [21].

2.4 Graphical Causal Models & Model Search

We adopt directed acyclic graphs (DAGs) to represent causal relationships among variables we seek to model. We consider the context of linear relations among variables and multi-variate Gaussian joint probability distributions, where DAGs imply conditional independence constraints on observed joint distributions and covariance matrices. The set of DAGs consistent with a set of independence constraints, assuming that there are no unmeasured common causes of measured variables, comprise an equivalence class of graphs, represented by a graphical object called a pattern. Patterns and other equivalence classes of graphs can be inferred from data by asymptotically reliable algorithms developed (e.g., the TETRAD[1] project) over the past 20+ years.

We deploy the constraint-based PC algorithm to learn a pattern from data, making the strong assumption of no unmeasured common causes of measured variables [22]. From a pattern, we can choose a DAG member of the equivalence class to specify a linear structural equation model (SEM). Allowing for unmeasured common causes, we consider an equivalence class of graphs, represented by Partial Ancestral Graphs (PAGs), learned using the FCI algorithm [22]. FCI is similar to PC, but PAGs have a richer set of edges between two variables X and Y in a PAG [13, 22]:

- X o—o Y: (1) X is an ancestor (i.e., cause) of Y; (2) Y is a cause of X; (3) X and Y share a latent common cause; (4) either (1) & (3) or (2) & (3).
- X o→ Y: Either X is a cause of Y; X and Y share a latent common cause; or both.
- $X \leftrightarrow Y$: X and Y share a latent common cause in every member of the equivalence class represented by this PAG.

[1] freely-available at http://www.phil.cmu.edu/projects/tetrad/

- $X \rightarrow Y$: X is an ancestor/cause of Y in every member of the equivalence class represented by this PAG.

3. DATA + PRIOR WORK

Our data are logs for a sample of 102 adult, higher education learners using CT Algebra. We consider log data over roughly 337,000 learner actions in a module of five units concerning linear equations and inequalities, relatively late in the course. We also have a pre-test score (*Module Pre-Test*) and a *Final Exam* score for the entire algebra course, which is our learning outcome.

Assuming no unmeasured common causes of variables, causal models of this data [13] illuminated one possible explanation for the positive correlations between both *Boredom* and *Confusion* and *Final Exam*: both may cause decreased *Gaming the System* behavior, behavior which is found to cause decreased learning. While *Confusion* may cause decreased *Gaming the System* (e.g., *Confusion* being an affective state in which learners are unlikely to be able to "game"), there are reasons to suspect that this correlation and others arise due to confounding common causes.

Relaxing the assumption of no unmeasured common causes and allowing affect and behavior to co-occur, the FCI algorithm found a *robust* causal link between gaming and learning; all other links in the PAG causal model from prior work are at least possibly confounded. This fact and several common cause hypotheses in the literature explaining positive links between *Confusion* and *Boredom* and learning lead us to consider *Carelessness*.

4. MODELING CARELESSNESS

4.1 Correlational Analysis

Carelessness is positively correlated with both *Module Pre-Test* ($r = 0.36$, $p < .001$) and *Final Exam* ($r = 0.56$, $p < .001$), consistent with results that careless behavior is common even among high-performing math learners [9-10]. Correlations of *Carelessness* to other affective and behavioral variables are presented in Table 1. These results are largely consistent with those in previous work analyzing the relationship between *Carelessness* and affect [21].

Table 1. Pairwise correlations of *Carelessness* and other variables representing "detected" behavior and affective states (*p < .05; ***p < .001)

Variable / Construct	Pearson Correlation
Boredom	0.13
Confusion	0.48***
Engaged Concentration	0.75***
Gaming the System	-0.74***
Off-Task Behavior	-0.25*

4.2 Causal Models

Rather than attempt to specify and test "by hand" a multitude of alternative models that posit different causal roles for *Carelessness*, we adopt a search strategy. Assuming that affective states (including *Carelessness*) causally precede behavioral variables, the PC algorithm learns the DAG causal structure of the estimated linear SEM of Figure 1. This model fits the data ($\chi^2(19) = 23$; $p = .22$) [7] and is similar to that the model found in previous work under the same assumptions [13]. We focus on three elements of it.

First, *Engaged Concentration* is inferred to be a cause of *Carelessness*, consistent with the high correlation in the Scatterplot Study [21], and hypotheses due to Clements [10] about the relationship between engagement (i.e., *Engaged Concentration*) and *Carelessness*. San Pedro, et al. note the positive link between confidence and *Carelessness* found by Clements and posit that an engaged learner of only average knowledge might become overly confident in their ability and careless [16, 21]. This explanation suggests an intermediary along this causal pathway, a topic for future research.

Second, *Carelessness* is inferred to be a common cause of *Confusion* and *Gaming the System*, with increased *Carelessness* leading to increased *Confusion* and less *Gaming the System*. *Carelessness* as a common cause of these two variables is consistent with models in [13] in which an edge *Confusion* $\circ\!\!\rightarrow$ *Gaming the System* indicated the possible presence of an unmeasured (i.e, confounding) common cause. The strong positive relationship between *Engaged Concentration* and the inferred cause of *Confusion*, *Carelessness*, provides a plausible explanation for the positive correlation of *Confusion* and learning, but this model does not suggest we pursue interventions that increase learner *Confusion*, though recent literature suggests that, in some contexts, *Confusion* may be beneficial for learning (e.g., [15]).

With respect to the other effect of *Carelessness* in Figure 1, *Gaming the System*, it is possible that there is a negative causal connection, as presumably gaming behavior is the result of at least a certain amount of non-careless affect and corresponding behavior, as learners must provide roughly appropriate responses to math problems if they are to, in fact, "game the system." However, it is also plausible that *Carelessness* and *Gaming the System* share a confounding common cause.

Relaxing the assumption of no unmeasured common causes and assuming only that *Module Pre-Test* precedes all affective and behavioral variables, all of which precede *Final Exam*, FCI learns the PAG causal model in Figure 2, with +/- signs to remind the reader of parameter estimates in Figure 1. Contrary to the model of Figure 1, either *Confusion* is a cause of *Carelessness*, or they share an unmeasured common cause. The direction of the link between *Confusion* and *Carelessness* is sensitive to the "ordering" of affective and behavioral variables. However, under nearly all combinations of behavioral and affective variable orderings and groupings, *Engaged Concentration* is a cause of *Carelessness*, consistent with past hypotheses [9-10] and correlational analyses [21]. While we infer that *Carelessness* and *Gaming the System* share an unmeasured common cause, relationships between variables like *Carelessness* and *Gaming the System* may be confounded, not only by other unmeasured phenomena, but by the underlying phenomenon itself since we provide only noisy measures using detector models.

5. DISCUSSION

We provide evidence for the hypothesis that concentration leads to (i.e., causes) careless mistakes, and this causal inference is robust under a variety of assumptions. Contrary to some hypotheses [8, 17], we do not find evidence for a causal link between *Carelessness* and *Boredom* in CT Algebra. However, that hypothesis of [17] was made with respect to the ASSISTments system. Future research should take on the problem of learning causal models from available observational data from systems like CT and ASSISTments to determine under what circumstances causal inferences of the sort we consider here generalize across

sub-populations using the same instructional system as well as different systems within the same (or different) domain.

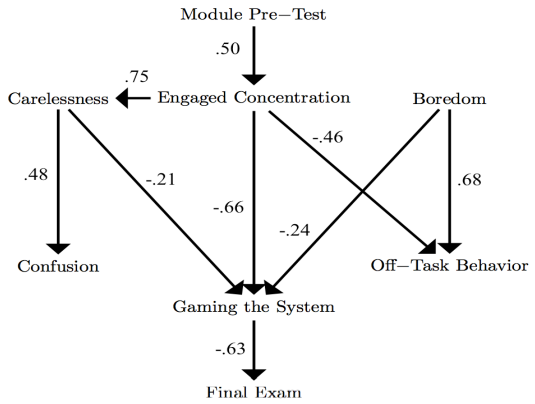

Figure 1. Estimated SEM incorporating *Carelessness*

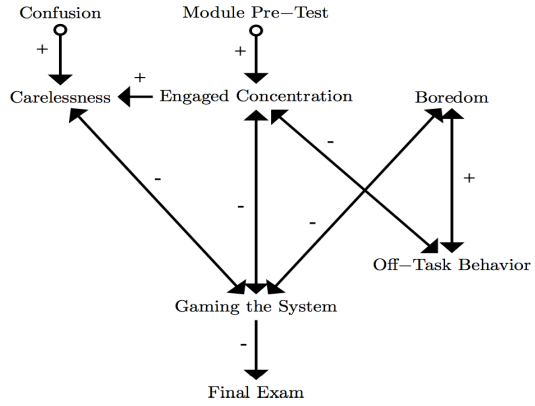

Figure 2. PAG causal model incorporating *Carelessness*

6. ACKNOWLEDGMENTS

The author gratefully acknowledges Ryan S.J.d. Baker, Susan R. Berman, Ryan Carlson, Sujith M. Gowda, Steven Ritter, and Charles Shoopak for providing code, assistance, and/or comments.

7. REFERENCES

[1] Baker, R.S.J.d. 2007. Modeling and understanding students' off-task behavior in intelligent tutoring systems. In *Proc. of ACM CHI 2007: Computer-Human Interaction* (San Jose, CA, April 28 – May 3, 2007). ACM, New York, 1059-1068.

[2] Baker, R.S.J.d., Corbett, A.T., Aleven, V. 2008. More accurate student modeling through contextual estimation of slip and guess probabilities in Bayesian Knowledge Tracing. In *Proc. of ITS 2008* (Montreal, Canada, 2008). 406-415.

[3] 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 Model. User-Adap.* 18 (2008), 287-314.

[4] Baker, R.S.J.d., de Carvalho, A. M. J. A. 2008. Labeling student behavior faster and more precisely with text replays. In *Proc. of EDM 2008* (Montreal, 2008). 38-47.

[5] Baker, R.S.J.d., Gowda, S.M., Wixon, M., Kalka, J., Wagner, A.Z., Salvi, A., Aleven, V., Kusbit, G.W., Ocumpaugh, J., Rossi, L. 2012. Towards sensor-free affect detection in Cognitive Tutor Algebra. In *Proc. of EDM 2012* (Chania, Greece, 2012). 126-133.

[6] Baker, R.S.J.d., Yacef, K. 2009. The state of educational data mining in 2009: a review and future visions. *Journal of Educational Data Mining* 1 (2009), 3-17.

[7] Bollen, K. 1989. *Structural Equations with Latent Variables*. John Wiley & Sons.

[8] Cheyne, J.A., Carriere, J.S., Smilek, D. 2006. Absent-mindedness: lapses of conscious awareness and everyday cognitive failures. *Conscious Cogn* 15 (2006), 578-592.

[9] Clements, M.A. 1980. Analyzing children's errors on written mathematical tasks. *Educational Studies in Mathematics* 11, (Feb. 1980), 1-21.

[10] Clements, M.A, 1982. Careless errors made by sixth-grade children on written mathematical tasks. *Journal for Research in Mathematics Education* 13, (Mar. 1982), 136-144.

[11] Cocea, M., Hershkovitz, A., Baker, R.S.J.d. 2009. The impact of off-task and gaming behavior on learning: immediate or aggregate? In *Proc. of AIED 2009* (Brighton, UK, 2009). 507-514.

[12] Corbett, A.T., Anderson, J.R. 1995. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Model. User-Adap.* 4 (1995), 253-278.

[13] Fancsali, S.E. 2014. Causal discovery with models: behavior, affect, and learning in Cognitive Tutor Algebra. In *Proc. of EDM 2014* (London, UK, 2014). 28-35.

[14] Feng, M., Heffernan, N.T., Koedinger, K.R. 2009. Addressing the assessment challenge in an intelligent tutoring system that tutors as it assesses. *User Model. User-Adap.* 19 (2009), 243-266.

[15] Lehman, B., D'Mello, S.K., Graesser, A.C. 2012. Confusion and complex learning during interactions with computer learning environments. *Internet High. Educ.* 15 (2012), 184-194.

[16] Linnenbrink, E.A., Pintrich, P.R. 2003. The role of self-efficacy beliefs in student engagement and learning in the classroom. *Reading & Writing Quarterly: Overcoming Learning Difficulties* 19 (2003), 119-137.

[17] Pardos, Z.A., Baker, R.S.J.d., San Pedro, M.O.C.Z., Gowda, S.M., Gowda, S.M. 2014. Affective states and state tests: investigating how affect and engagement during the school year predict end-of-year learning outcomes. *Journal of Learning Analytics* 1 (2014), 107-128.

[18] Pekrun, R., Goetz, T., Titz, W., Perry, R.P. 2002. Academic emotions in students' self-regulated learning and achievement: a program of quantitative and qualitative research. *Educ. Psychol.* 37 (2002), 91-106.

[19] Ritter, S., Anderson, J.R., Koedinger, K.R., Corbett, A.T. 2007. Cognitive Tutor: applied research in mathematics education. *Psychon. B. Rev.* 14 (2007), 249-255.

[20] San Pedro, M.O.C.Z., Baker, R. S. J. d., Rodrigo, M. 2011. Detecting carelessness through contextual estimation of slip probabilities among students using an intelligent tutor for mathematics. In *Proc. of AIED 2011* (Auckland, New Zealand). 304-311.

[21] San Pedro, M.O.C.Z., Baker, R.S.J.d., Rodrigo, M.T. 2014. Carelessness and affect in an intelligent tutoring system for mathematics. *Int J Artif Intell Educ* 24 (2014), 189-210.

[22] Spirtes, P., Glymour, C., Scheines, R. 2000. *Causation, Prediction, and Search*. 2nd Edition. MIT, Cambridge, MA.