Causal Discovery with Models: Behavior, Affect, and Learning in Cognitive Tutor Algebra

Stephen E. Fancsali
Carnegie Learning, Inc.
437 Grant Street, Suite 918
Pittsburgh, PA 15219 USA
1.888.751.7094 x219
sfancsali@carnegielearning.com

ABSTRACT
Non-cognitive and behavioral phenomena, including gaming the system, off-task behavior, and affect, have proven to be important for understanding student learning outcomes. The nature of these phenomena requires investigations into their causal structure. For example, given that gaming the system has been associated with poorer learning outcomes, would reducing such behavior improve outcomes? Answering this question requires an understanding of whether gaming the system is a cause of poor outcomes, rather than, for example, only sharing a common cause with factors influencing learning. Because controlled experiments to settle such causal questions are often costly or impractical, we employ algorithmic search for the structure of graphical causal models from non-experimental data. Using sensor-free, data-driven detectors of behavior and affect, this work extends Baker and Yacef’s notion of “discovery with models” to incorporate causal discovery and reasoning, resulting in an approach we call “causal discovery with models.” We explore a case study of this approach using data from Carnegie Learning’s Cognitive Tutor for Algebra and raise questions for future research.

Keywords
Discovery with models, causal discovery, graphical causal models, probabilistic graphical models, gaming the system, affect, off-task behavior, sensor-free detectors, intelligent tutoring systems, Cognitive Tutor, measurement.

1. INTRODUCTION
Recently, researchers in educational data mining, learning analytics, and the learning sciences have used the moniker “discovery with models” to describe analyses in which “a model of a phenomenon is developed through any process that can be validated in some fashion… and this model is then used as a component in another analysis, such as prediction or relationship mining” [10]. Examples of discovery with models range over a variety of constructs that capture student context and interaction with educational software and courseware [22] like help-seeking strategies [2] and patterns of use of online resources (e.g., [23]).

We focus on models that function as sensor-free “detectors” that use data from student interactions with an intelligent tutoring systems (ITS) to predict whether actions are likely instances of particular forms of behavior or arise from a student being in a particular affective state. Such detectors (and corresponding constructs of interest) have been the topic of a great deal of literature in educational data mining and the learning sciences; predicted constructs include “gaming the system” [6,8], off-task behavior [3], affective states (e.g., boredom and engaged concentration) [9], and carelessness [39], among others. These detectors are generally validated by comparing their data-driven predictions (e.g., whether a student is likely to be gaming the system or bored at a particular [interval of] time) to classifications provided by trained observers in a classroom or computer laboratory environment (cf. [30]).

While discovery with models approaches have been used to associate learning outcomes with various constructs, we suggest that many constructs of recent interest are especially important because of underlying causal questions: What causes behavior like gaming the system? What makes students bored, careless, or frustrated? What are causal links (if any) among such modeled constructs, and are they causally linked to outcomes like learning? Gaming the system, for example, and learning are found to be negatively associated in several studies (e.g., [14,31]), but the mere association of gaming behavior and learning does not imply that if we reduced gaming we would increase learning. Perhaps both are caused by some other factor (like motivation), making a focus on gaming behavior itself ineffective in increasing learning. Ideally, researchers would settle causal questions using randomized experiments (e.g., A/B software tests, randomized controlled trials), but often such experiments, if possible, are expensive, difficult, or unethical. Given non-experimental ITS log data, and its wide availability from sources like the Pittsburgh Science of Learning Center’s DataShop [26], we turn to algorithmic methods to discover causal structure from observational data.

After describing Carnegie Learning’s Cognitive Tutor® (CT) Algebra [36] ITS and several constructs for which sensor-free detectors have been developed, we briefly explicate the framework of data-driven search for the structure of graphical causal models. We then apply this framework in a discovery with models approach to find causal explanations that integrate aspects of behavior, affect, and learning in ITSs. Finally, we describe several important and interesting problems, especially but not limited to measurement problems, at the intersection of causal discovery from non-experimental data and discovery with models in educational data mining (i.e., “causal discovery with models”).

2. PRELIMINARIES

2.1 Cognitive Tutor (CT) Algebra
CT Algebra is an ITS with hundreds of thousands of middle school and high school users both in the United States and internationally. Increasingly, CT Algebra is also deployed in higher education settings. The CT adaptively presents
mathematics content to students by tracking their mastery of fine-grained knowledge components (KCs) or skills, into which mathematics content has been atomized, as they work through (parts of) problems (cf. the screenshot of Figure 1). At each problem-solving step, students can request context-sensitive hints and receive immediate feedback about correctness that is sometimes accompanied by just-in-time, context-sensitive feedback that is more detailed.

![Figure 1. Screenshot of problem solving in CT Algebra](image)

The CT deploys content that is divided into curricular units comprised of (roughly topical) sections. When the CT judges a student, using a framework called Bayesian Knowledge Tracing (BKT) [15], to have reached mastery of all the KCs in a particular section, she is graduated to the following section (or unit if she has completed all the sections in a given unit).

### 2.2 Data-Driven Detectors of Behavior and Affect

Recent work has developed a variety of data-driven, sensor-free “detectors” to infer or measure different features of student interactions with educational courseware, especially ITSs. In this work, we focus on using detectors to infer aspects of learners’ gaming the system, off-task behavior, and affective states while interacting with the CT Algebra ITS.

#### 2.2.1 Gaming the System & Off-Task Behavior

A great deal of recent work has been directed at using data-driven, predictive models to measure or infer disengaged behavior, including gaming the system and off-task behavior, and linking such behavior to learning outcomes using discovery with models techniques. Gaming the system [5-7] is characterized as behavior that allows for progression through curricular material without genuine learning by taking advantage of ITS affordances available to the learner. In general, such behavior can be broadly characterized by learners’ abuse of hints [1], including cycling through hints until the last hint (i.e., a “bottom out” hint) is reached that provides the answer to a problem-solving step, and by rapid and/or systematic guessing [7]. While some gaming behavior has been called “non-harmful” because it is not associated with decreased learning, there is a great deal of evidence for “harmful gaming” that is associated with decreased learning [6,7]. The “harmful” modifier can (at least tacitly) be read causally, even if generally used to describe merely correlational results, so one research question involves determining whether we can provide evidence from non-experimental data for the claim that gaming the system is a *cause of decreased learning*. Off-task behavior refers to behavior that is disengaged and/or unrelated to learning or the learning environment [3].

A variety of data-driven detectors of gaming the system and off-task behavior have been developed in recent years (e.g., [11,24,43]). We deploy detectors of gaming the system [8] and off-task behavior [3] developed for CT Algebra, the statistical basis of which are Latent Response Models [29]. These detectors employ features “distilled” from fine-grained CT process data that capture the types of behavior we described above (e.g., for gaming the system: quick actions after making at least one error on a problem-solving step [7]; for off-task behavior: extremely fast or extremely slow actions [3]). Detectors generate a prediction for each learner action in the CT as to whether it is likely an instance of gaming the system or off-task behavior. These predictions can then be “rolled-up” to the level of consecutive actions on a particular problem solving-step (i.e., roughly consecutive actions involving the same KC). If any one action within a problem-solving step is determined to be gamed or off-task, then we call that step gamed or off-task, following other applications of these detectors (e.g., [14]).

#### 2.2.2 Affective States

While evidence suggests that learner affect can influence learning (e.g., [16,33]), measurement and assessment of affect, whether via surveys, physical sensors, or direct observation, can be obtrusive, time-consuming, and suffers from a lack of scalability to larger numbers of learners over longer periods of time. In an effort to overcome these obstacles, recent work [9] has taken a data-driven, sensor-free approach to infer learner affect from ITS process data, much like that adopted to infer gaming and off-task behavior.

As with gaming and off-task detectors, in the development of affect detectors, a wide variety of features are distilled from CT process data, but rather than learning a Latent Response Model, machine learning classifiers are applied to features of “clips” of problem-solving (durations of learner actions up to twenty seconds in length) to classify them as likely corresponding to learners being in a state of boredom, confusion, engaged concentration, or frustration. Boredom in a particular clip can be detected, in part, through features like the maximum number of previous incorrect actions and hint requests for any skill in the clip. Confusion in a clip can be detected, for example, using the percentage of actions that take longer than five seconds after two incorrect answers. The duration of the fastest action in a clip is one feature upon which the detector of engaged concentration relies. Finally, frustration can be detected, in part, by the percentage of past actions on skills in a clip that were incorrect [9].

These classifiers can then be applied to new data to generate predictions about problem-solving clips and their correspondence to learner affective states. We now review several successful instances of discovery with models approaches using detectors to predict external, student-level learning outcomes. These results demonstrate correlations between learning outcomes and gaming the system, off-task behavior, and affective states, but we seek further insight into whether non-experimental data alone can provide evidence for causal claims about the impact of these phenomena on learning.

### 2.3 Prior Work: Using Models of Behavior & Affect to Predict Learning Outcomes

Several recent projects have used data, aggregated over fine-grained predictions, from these detectors as input to statistical,
predictive analyses of student-level, substantive learning outcomes (i.e., adopting a discovery with models approach). One study used aggregate counts of gaming and off-task problem-solving steps to predict post-test scores for several units of CT content [14]. These researchers built linear regression models for each CT unit they considered, summarizing their results by reporting that gaming the system was weakly associated, and off-task behavior strongly associated, with poorer learning in the aggregate.

Later work has successfully used detectors of gaming the system, off-task behavior, and affect on data from the ASSISTments system [21], to predict Massachusetts Comprehensive Assessment System (MCAS) standardized test scores [31] and college enrollment (in a different population and study) some time after using the software [38]. The former study reports pairwise correlations between variables aggregated from detector predictions and raw MCAS scores, treating different types of ASSISTments’ problems separately. For our purposes, ASSISTments’ “scaffold” problems, presented after a student has made a mistake or asked for help, are most relevant to understanding student behavior in the CT, as their structure is the norm for CT problems.

Boredom, confusion, engaged concentration, and frustration on scaffold problems are all positively and significantly correlated with MCAS scores across two academic years of data. They report mixed results (one year positive, one negative) about the correlation of off-task behavior with MCAS scores and that gaming the system is significantly, negatively correlated with MCAS scores. A logistic regression model of college enrollment based on detectors also found significant, positive associations between enrollment and both boredom and confusion [38].

Having summarized several correlational studies that exemplify discovery with models using data-driven detectors, we now introduce the framework of graphical causal models and algorithmic search procedures to learn causal structure from non-experimental data. Our aim will then be to use detector predictions as input to these procedures to go beyond analysis of correlations.

3. GRAPHICAL CAUSAL MODELS & CAUSAL DISCOVERY

To learn causal relationships among the phenomena of gaming the system, off-task behavior, affective states, and learning, we adopt the formalism of graphical causal models, specifically causally interpreted directed acyclic graphs (DAGs), to represent the qualitative causal structure among variables of interest [32,42]. Such models have been used to better understand causal relationships among various phenomena in ITSs (e.g., [20,34,35]) and in educational technology more generally (e.g., [17,40]).

Under the causal interpretation of DAGs, nodes represent random variables and directed edges represent direct causal relationships, relative to the set of variables in the model. For linear causal relations and multivariate normal joint probability distributions, DAGs imply (conditional) independence constraints on observed distributions or covariance matrices. Whether particular constraints obtain for observed data can be ascertained by statistical tests for whether appropriate partial correlations vanish.

However, it is often the case that more than one DAG is consistent with the same set of (conditional) independence constraints; that is, causal structure is underdetermined by non-experimental data, and members of the set of all of DAGs consistent with the same constraints (i.e., members of an equivalence class of DAGs) are indistinguishable from observation alone. Consider, for example, a simple case of three observed variables X, Y, and Z, no pair of which shares any unmeasured common cause. Suppose the pair-wise correlation of each pair of variables is non-zero, and that by a statistical test we determine that the sample partial correlation \( \rho_{XY|Z} \) vanishes (i.e., \( \rho_{XY|Z} = 0 \)). Three DAGs are consistent with this conditional independence relationship (i.e., are members of the equivalence class consistent with this constraint):

- \( X \rightarrow Z \rightarrow Y \)
- \( X \leftarrow Z \rightarrow Y \)
- \( X \leftarrow Z \leftarrow Y \)

Beyond data-driven constraints, background and domain knowledge are also important. If we knew, for example, that Z were prior in time to X and Y, then only one graph (the graph in which Z is a common cause of X and Y: \( X \leftarrow Z \rightarrow Y \)) is consistent with both the conditional independence constraint and background knowledge.

Researchers have developed asymptotically reliable algorithms1 [42] to infer the equivalence class of causal graphs that are consistent with observed independencies, conditional independencies, and background knowledge, under different assumptions. We focus on two such algorithms. The PC algorithm [42] learns a graphical object called a pattern that represents the equivalence class of DAGs consistent with observed (conditional) independencies and background knowledge, assuming there are no unmeasured (i.e., latent) common causes of measured variables. Qualitative causal structure of a DAG member of the class represented by a pattern (each member of which will fit the data equally well) can be used to specify a linear structural equation model. Estimating parameters of such a model allows for path analysis and the consideration of quantitative causal effects (as we will see in §5.3.1).

Since the assumption of no latent common causes is implausible for most real-world scientific settings, we also consider search using the FCI algorithm [42], which allows for the possibility of latent common causes. While FCI is similar to PC in many ways, the graphical object it learns from data, called a Partial Ancestral Graph (PAG), also represents an equivalence class of causal graphs but is more expressive to allow for possible latent common causes. Edges in PAG causal models are interpreted as follows:

- \( X \perp \! \! \! \perp Y \) (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).
- \( X \perp \! \! \! \perp Y \) (i.e., X is a cause of Y; and Y share a latent common cause; or both.
- \( X \leftrightarrow Y \) (i.e., X and Y share a latent common cause in every member of the equivalence class represented by this PAG.
- \( X \rightarrow Y \) (i.e., X is an ancestor/cause of Y in every member of the equivalence class represented by this PAG.

A graph containing this last type of edge represents a case where we can make causal inferences despite the assumption that there may be latent common causes of measured variables. We now summarize our data as well as how we construct variables, from

---

1 implemented and made freely-available by the Tetrad Project (http://www.phil.cmu.edu/projects/tetrad/)
predictions of detector models, to use as input to causal structure search algorithms.

4. DATA

4.1 Overview

We consider CT Algebra log data over a sample of 102 learners who completed an algebra course in a higher education context. Specifically, we focus on a module of CT Algebra units presented at the end of a particular course that included the following units of instruction:

- Systems of Linear Equations
- Systems of Linear Equations Modeling
- Linear Inequalities
- Graphing Linear Inequalities
- Systems of Linear Inequalities

In addition to pre-test scores for this module of instruction and final exam scores for the entire algebra course for each of the 102 learners, we constructed aforementioned data-driven detectors of gaming the system, off-task behavior, and various affective states, including boredom, confusion, engaged concentration, and frustration, from fine-grained log files containing roughly 337,000 student actions. We learned BKT parameters, required as input to these detectors, for the 32 KCs in our data using a brute-force method [4].

4.2 Variable Construction

Since it is implausible that any particular interaction (e.g., gaming the system on a particular problem solving step in CT) is attributable as a cause of aggregate student learning, we seek aggregate patterns of interaction (i.e., variables aggregated at the level of students) over which to learn causal models and provide causal explanations. That is, our present project is to provide causal models that could explain relationships among student-level behavior, affect, and learning, but the results of the detector models we seek to use as a component of (i.e., as input to) causal search algorithms are fine-grained (i.e., predictions about behavior and affect during many problem solving steps or clips of interaction per student).

Previous work [18,19] provided a preliminary exploration of this data set using algorithmic causal search over variables defined as student-level counts of gamed or off-task problem-solving steps (as assessed by appropriate detectors), following other work that modeled aggregate variables constructed from predictions of these detectors [14]. The current project extends this exploration by integrating detectors of learner affect and constructs variables differently, roughly following more recent, aforementioned research using detectors to predict learners’ MCAS scores [31].

We define variables for the proportion of problem-solving steps, per learner, that are judged to be instances of gaming and off-task behavior and the proportion of problem-solving clips (i.e., longer durations of problem-solving activity) at which learners are judged to be in particular affective states. Constructing variables in this way allows for each variable to represent the proportion of the student’s CT interaction in which they behaved in a particular way or were inferred to be in a particular affective state, eliminating the complication of counting steps versus clips for the two different types of detectors deployed. Despite modest differences in variable construction, high-level results we now present are consistent with these previous modeling efforts using variables defined or constructed as counts.

5. RESULTS

We begin describing our results by summarizing learner behavior and affect. Then we present pair-wise correlations of modeled behavior and affect variables with learning before presenting structural, causal models that explain patterns of conditional independence among these measures.

5.1 Relative Frequencies of Behaviors and Affective States

As a check on the applicability of the detectors, we consider the relative frequency with which particular behavioral and affective predictions are made by the detectors we used. Our findings roughly align with previous applications to data from (and observations of) CT Algebra and other tutors (cf. [9]; [S.M. Gowda, personal communication]). Over all 102 students and all usage, 41% of steps are determined to be instances of gaming the system while 4.4% of steps are deemed off-task. The detectors of affect infer that 5.87% of all problem-solving clips correspond to learner boredom; 3.52% of clips correspond to learner confusion; 67.5% of clips are inferred to be instances of engaged concentration, and 8.6% of clips correspond to frustration.

The relatively large percentage of gaming the system may be partially attributable to the fact that the detector, as in previous studies of the aggregate impact of gaming (e.g., [14]), does not distinguish between what has been called “harmful” and “non-harmful” gaming. Moreover, some behavior inferred to be gaming might be helpful for learning, as, for example, when students seek “bottom out” hints as worked examples [41].

5.2 Correlations with Learning

Next, we consider pairwise correlations of each modeled behavior and affective state variable and our learning outcome, the algebra course final exam score. We report Pearson correlation coefficients in Table 1 (noting significance of each according to the two-tailed t-test for such coefficients).

It is perhaps unsurprising that Gaming the System, Off-Task Behavior, and Frustration are negatively correlated with learning and Engaged Concentration is positively correlated with learning. That Boredom and Confusion are both positively correlated with learning, while perhaps surprising, is consistent with predictive results reported in both the MCAS and college enrollment studies we briefly summarized in §2.3 that used ASSISTments data. In considering causal models of these constructs in the following section, we provide possible explanations for the directions of these correlations and associations.

<table>
<thead>
<tr>
<th>Variable / Construct</th>
<th>Pearson Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boredom</td>
<td>.18</td>
</tr>
<tr>
<td>Confusion</td>
<td>.31**</td>
</tr>
<tr>
<td>Engaged Concentration</td>
<td>.55***</td>
</tr>
<tr>
<td>Frustration</td>
<td>-.27**</td>
</tr>
<tr>
<td>Gaming the System</td>
<td>-.63***</td>
</tr>
<tr>
<td>Off-Task Behavior</td>
<td>-.09</td>
</tr>
</tbody>
</table>

5.3 Causal Model Discovery

To learn structural, causal models to help explain these pairwise correlations, we apply aforementioned search algorithms to learn qualitative causal structure. So that we might provide a robust analysis, we consider three different sets of assumptions, including the temporal ordering of variables as background...
knowledge, that constrain the search for causal structure and the types of inferences that can be made. The issue of temporal ordering is especially important given a lack of agreement about whether behavior precedes affect, vice versa, or they co-occur (cf. [9]). We begin with the strongest assumptions, relax those assumptions, and then briefly consider robustness of causal inferences across these assumptions.

5.3.1 No Unmeasured Common Causes & Affect Precedes Behavior

We begin with two relatively strong assumptions. First, we assume that there are no unmeasured common causes of measured variables. This assumption is unlikely to hold in most real world settings. Second, we assume that learner affect causally precedes behavior. We constrain the PC algorithm's search space by providing background knowledge that Module Pre-Test precedes Confusion, Engaged Concentration, and Boredom, and that these three affective states precede Gaming the System and Off-Task Behavior. Finally, the course Final Exam is the last variable in this ordering.2

Using the qualitative causal structure inferred with the PC algorithm,3 we specify and estimate parameters of the linear structural equation model (graphically represented) in Figure 2. In such a model, each variable is a linear function of its parents (i.e., direct causes) and an independent, normally distributed error term (omitted from Figure 2). This linear model fits the observed data well, as assessed by a chi-square test comparing the implied covariance matrix to the observed covariance matrix ($\chi^2(13) = 14.64, p = .33$) [13]. While we will find that the inferred causal relationship between Gaming the System and Final Exam (i.e., learning) is robust across all sets of assumptions we will consider, given the assumptions we have made so far, other edges in the graph should be interpreted cautiously.

Keeping this caveat in mind, considering the structure and parameters of the model in Figure 2, we see that Module Pre-Test is directly linked only to Engaged Concentration; learners with greater pre-test scores tend to concentrate more. Learners with a higher proportion of Engaged Concentration tend to game the system less and go off-task less, and we have already seen that gaming is strongly, negatively correlated with Final Exam, our learning outcome.

Off-Task Behavior has no direct effect on Final Exam (or on Gaming the System), which we will see is also a robust finding. Interestingly, since Confusion is positively correlated with Engaged Concentration and negatively correlated with Gaming the System, one explanation of the positive, pair-wise correlation of Confusion and Final Exam is that increased Confusion virtually leads both directly and indirectly (via leading students to better concentration) to less Gaming the System. This provides one possible causal explanation consistent with recent literature showing that confusion can be beneficial for learning (e.g., [27,28]).

![Figure 2. Estimated linear structural equation model](image)

That increased Boredom may contribute to less Gaming the System and better learning is consistent with an aforementioned finding [31], but the proposed explanation in that work posits an unmeasured common cause (here also unmeasured) of boredom in “scaffold” questions and better learning. Careless, for example, rather than a lack of skill mastery, may drive students to answer incorrectly on originally presented questions, forcing learners into ASSISTments’ scaffold questions; consequently learners become bored. Apropos, we consider relaxing the assumption that there are no unmeasured common causes and of the ordering of affect and behavior, allowing that neither precedes the other, but rather that they may co-occur.4

5.3.2 Relaxing Assumptions

The result of relaxing these two assumptions and applying the FCI algorithm to our data is the PAG causal model of Figure 3. First, despite relaxing both of our relatively strong assumptions, we still make the positive inference that Gaming the System is a cause of decreased learning (i.e., is generally “harmful” in the aggregate). Next, we see that inferring affective causes of Gaming the System is more complicated. Both Engaged Concentration and Boredom are found to share at least one unmeasured common cause with Gaming the System; the same is true of their relationships with Off-Task Behavior. However, despite relaxing the ordering of affect preceding behavior, we still find that Confusion is possibly a cause of Gaming the System, though the two may also share an unmeasured common cause. Finally, Module Pre-Test and Engaged Concentration are also possibly confounded, but this is perhaps unsurprising because, at best, such a pre-test is a noisy measure of prior ability.

---

2 We omit Frustration from our analysis because it is relatively rare, and we were unsuccessful in inferring linear models that fit observed data well when we included it in the search. Future work should determine whether frustration in ITS environments is so rare. Assuming we are inferring a valid affective feature, finding appropriate means for analysis is also an important topic for future work.

3 While in general we learn an equivalence class of causal graphs with the PC algorithm, in this case, our assumption of temporal ordering leads us to the unique DAG structure illustrated by Figure 2.

4 While perhaps less likely given the relatively short span of the single CT Algebra module we consider, cyclic relationships over time between behavior and affect might be fruitfully treated in an acyclic setting by constructing appropriate, time-indexed (e.g., section-by-section or unit-by-unit) aggregate variables. This is a topic for future research.
In the case in which affect precedes behavior (as we assume in the methods for causal discovery from non-experimental data, the System cede affect, positive causal inferences are also made that demonstrated that this approach can be used to provide evidence about relationships among important constructs of interest that are confusion),

Exam Gaming the System

Brainly trivially lead to less Confusion (i.e., attempts to avoid genuinely learning should not increase confusion), Gaming the System could both impinge upon concentration and decrease Boredom in more substantive ways. In the case in which affect precedes behavior (as we assume in the model of Figure 2), we find that Engaged Concentration causes decreases in both Off-Task Behavior and Gaming the System, but relationships between Confusion, Boredom, and Gaming the System may be confounded.

Simply because more positive causal inferences can be made given a particular temporal ordering does not provide evidence that we have arrived at the “correct” ordering. Questions like this should be settled by some combination of theoretical considerations, experimental results, and data analysis. The most important conclusion we reach from examining different sets of assumptions is that the positive inference about Gaming the System (i.e., that it is harmful to learning, in the aggregate) is robust across all of them. The negative finding that Off-Task Behavior is not a cause of learning is also robust. Lacking robustness for other inferences, we find the model of Figure 3 with weaker (i.e., modest) background knowledge, allowing that behavior and affect co-occur, most plausible and return to it in our discussion.

6. DISCUSSION

This work makes at least two important contributions. First, we have illustrated an approach, combining discovery with models and methods for causal discovery from non-experimental data, that we call causal discovery with models. Second, we have demonstrated that this approach can be used to provide evidence about relationships among important constructs of interest that are not always practical targets of randomized experiments. Specifically, we have provided evidence that gaming the system is, in fact, harmful (i.e., both negatively correlated and likely causally related) to aggregate learning for a relatively novel sample of learners in a higher education context.

Notably, we do not have ground truth labels for our data (e.g., field observations of whether students are off-task or bored), so in this sense this work helps to generalize the idea that these detectors can be applied to data from new student populations and yield interesting connections to learning. Nevertheless, further research is necessary to determine what unmeasured common causes may confound relationships between affect and behavior, either because of measurement problems or because of phenomena we have not included in these models (e.g., carelessness, motivation, etc.).

A perhaps underappreciated problem for a variety of discovery with models approaches concerns the process by which the output or results of particular models (here, fine-grained detector predictions concerning behavior and affect) are used to construct variables that are used as input in other analyses (here, causal graph search). How do we best use the output or results of a particular model as a component (possibly components) of other analyses? Previous work [18,19] explored this problem as one of semi-automated search for constructed variables (including several different levels of aggregation and aggregation functions as suggested by [14], but future work should explore better aggregate variable construction and feature engineering (including considerations of the interpretability of resulting features) as well as measurement models for these constructs.

To wit, despite sophisticated feature engineering used by detectors to make predictions and classifications, variables included in these models provide a relative paucity of information about the underlying phenomena of interest. More sophisticated measurement models might be used to either explicitly model latent phenomena or to develop improved, measured proxies (e.g., scales) to represent these constructs. As operationalized, Boredom and Off-Task Behavior may, for example, be confounded by boredom itself, as the constructed variable Boredom is only one noisy measure of the underlying phenomenon.

We should also include more phenomena in these models, including both latent phenomena like motivation and learner goals [12,20] and relatively simple process measures from CT log data. Prior work with this data set [18,19], for example, found that the count of actions that trigger context-sensitive, just-in-time feedback in the CT, while possibly less tenable as a target for future interventions, is highly correlated both with the final exam score and gaming the system. These prior results also suggest that this measure, or more likely the phenomena for which it stands in as a proxy, is an intermediate link in a causal chain from Gaming the System to learning. Such measures, including other relatively simple measures of learner efficiency and assistance required by learners (e.g., the “assistance score” that sums hints requested and errors made [26]) that have been found to predict standardized test

---

5 This is not to say that phenomena like gaming the system have not or cannot be targets of interventions (cf. [3]). However, the design and implementation of many experiments is likely to be non-trivial.

6 possibly shallow learning or a learner’s tendency to simply enter values that appear in a problem.
scores for middle school CT users [37], do not require sophisticated feature engineering to achieve predictive access to learning outcomes and have been found to be highly correlated with gaming the system [25]. These and other measures should be further evaluated and explored. Their suitability to preserve (or induce) appropriate conditional independence relationships, necessary for modeling causal relationships in the framework we have described (given assumptions we have considered), should also be evaluated.

7. ACKNOWLEDGMENTS
Ryan S.J.d. Baker, Sujith Gowda, and Charles Shoopak provided code for (and vital assistance with) building predictive, “detector” models for gaming the system, off-task behavior, and student affect. I also thank Susan Berman, Ryan Carlson, David Danks, Ambarish Joshi, Tristan Nixon, Steven Ritter, Richard Scheines, and Michael Yudelson for their insightful comments and suggestions as this work has developed.

8. REFERENCES


