Discovery with Models:

A Case Study on Carelessness in Computer-based Science Inquiry

Arnon Hershkovitz, Ryan S.J.d. Baker

Department of Human Development, Teachers College Columbia University

Janice Gobert, Michael Wixon, Michael Sao Pedro

Department of Social Sciences and Policy Studies, Worcester Polytechnic Institute

Hershkovitz, A., Baker, R.S.J.d., Gobert, J., Wixon, M., Sao Pedro, M. (2013). Discovery with models: A case study on carelessness in computer-based science inquiry. *American Behavioral Scientist, 57* (10), 1479-1498.

Abstract

In recent years, an increasing number of analyses in Learning Analytics and Educational Data Mining (EDM) have adopted a "Discovery with Models" approach, where an existing model is used as a key component in a new EDM/analytics analysis. This article presents a theoretical discussion on the emergence of discovery with models, its potential to enhance research on learning and learners, and key lessons learned in how discovery with models can be conducted validly and effectively. We illustrate these issues through discussion of a case study where discovery with models was used to investigate a form of disengaged behavior, i.e., carelessness, in the context of middle school computer-based science inquiry. This behavior has been acknowledged as a problem in education as early as the 1920s. With the increasing use of high-stakes testing, the cost of student carelessness can be higher. For instance, within computer-based learning environments careless errors can result in reduced educational effectiveness, with students continuing to receive material they have already mastered. Despite the importance of this problem, it has received minimal research attention, in part due to difficulties in operationalizing carelessness as a construct. Building from theory on carelessness and a Bayesian framework for knowledge modeling, we use machine-learned detectors to predict carelessness within authentic use of a computer-based learning environment. We then use a discovery with models approach to link these validated carelessness measures to survey data, to study the correlations between the prevalence of carelessness and student goal orientation.

Keywords: discovery with models, learning analytics, educational data mining, carelessness, science inquiry, goal orientation

Discovery with Models:

A Case Study on Carelessness in Computer-based Science Inquiry

The recent explosion of student interaction data derived from educational software, has resulted in a proliferation of models developed through Educational Data Mining (Baker & Yacef, 2009; Romero & Ventura, 2007, 2010) and Learning Analytics (Siemens & Long, 2011) techniques. These relatively new fields provide analytic techniques to operationalize a range of learner behaviors, attributes, and states during the learning process. In recent years, these models have proven to be useful for providing formative data to instructors, and for the design and implementation of automated learning interventions (cf. Arnold, 2010; Arroyo, Woolf, Cooper, Burleson, & Muldner, 2011; Militello & Heffernan, 2009). Increasingly, the establishment of these models has also become a powerful tool for making scientific discoveries about learning and learners, through studying the contexts in which a learning-related behavior occurs or construct emerges, and through examining its relationships with other constructs. Baker and Yacef (2009) termed the use of an existing EDM/analytics model as a component in a new EDM/analytics analysis "discovery with models", and noted its increasing prevalence in EDM research. Discovery with models has been used to study a range of constructs, including help-seeking strategies (Aleven, Mclaren, Roll, & Koedinger, 2006), off-task behavior (Baker, 2007b; Baker & Gowda, 2010), patterns of usage of online resources (Jeong & Biswas, 2008; Kinnebrew, Biswas, & Sulcer, 2010), and social networks (Dawson, Macfadyen, Lockyer, & Mazzochi-Jones, 2011).

In this paper, we illustrate discovery with models methods drawing on examples involving two constructs: gaming the system (Baker et al., 2004; Baker, Walonoski,

Heffernan, Roll, Corbett, & Koedinger, 2008b; Baker, 2007a; Baker & Gowda, 2010; Muldner, Burleson, Van de Sande, & VanLehn, 2011) and carelessness (cf. Georges, 1929; McClure, 1929). Gaming the system is a student's behavior aiming at succeeding in an educational task by systematically taking advantage of properties and regularities in the system used to complete that task, rather than by thinking through the material (e.g., by systematically guessesing answers in order to obtain the correct response, or repeatedly requesting hints until the software presents the answer; Baker et al., 2004). The second construct, carelessness, refers to incorrect answers given by a student on material that the student should be able to answer correctly (Newman, 1977). Careless errors are possibly due to answering impulsively or with insufficient effort (cf. Rodriguez-Fornells & Maydeu-Olivares, 2000).

The application of discovery with models involves two main phases. First, a model of a construct is developed using machine learning or knowledge engineering techniques, and is then validated, as discussed below. Second, this validated model is applied to data and used as a component in another analysis: For example, for identifying outliers through model predictions; examining which variables best predict the modeled construct; finding relationships between the construct and other variables using correlations, predictions, associations rules, causal relationships or other methods; or studying the contexts where the construct occurs, including its prevalence across domains, systems, or populations. For example, in Baker et al. (2008b), data from a validated machine-learned model of gaming the system was applied to data from an online learning system and used in a correlational analysis, in combination with survey

data, to determine which of a range of motivational and attitudinal variables were associated with gaming.

One essential question to pose prior to a discovery with model analysis is whether the model adopted is valid, both overall, and for the specific situation in which it is being used. Ideally, a model should be validated using an approach such as cross-validation, where the model is repeatedly trained on one portion of the data and tested on a different portion, with model predictions compared to appropriate external measures, for example assessments made by humans with acceptably high inter-rater reliability, such as field observations of student behavior for gaming the system (cf. Baker, Corbett, & Koedinger, 2004; Baker et al., 2008a), or future student performance. Validating using measures external to the model itself increases the confidence that the model is genuinely assessing what it was intended to measure. By testing generalizability to unseen data, one can ensure that a model will be usable in new contexts. When testing a model's generalizability using an approach such as cross-validation, it is important to divide the data into training and test folds at the appropriate level. For example, to ensure that a model can be validly applied to new students, the model should be cross-validated at the student level, e.g. each student appears either in the training or test set at a specific time. Even after validating in this fashion, validity should be re-considered if the model is used for a substantially different population or context than was used when developing the model..

An alternative approach is to use a simpler knowledge-engineered definition, rationally deriving a function/rule that is then applied to the data. In this case, the model can be inferred to have face validity. However, knowledge-engineered models often

produce different results than machine learning-based models, for example in the case of gaming the system. Research studying whether student or content is a better predictor of gaming the system identified different results, depending on which model was applied (cf. Baker, 2007a; Gong et al., 2010; Muldner et al., 2011). Of course, knowledge-engineered and machine-learned models do not always disagree; for instance, both types of models have shown similar relationships between gaming and learning or motivation/attitude (Aleven et al., 2006; Baker et al., 2008b; Cocea et al., 2009). When differences emerge, it is often because knowledge-engineered models and machine-learned models do not quite predict the same constructs: there is a trade-off between the knowledge-engineered models' clarity, and the machine learning models' precise prediction which is based on finding complex and unexpected relationships.

Discovery with models has several advantages as a method. First, it allows for scientific investigation to be made using definitions that can be discussed and inspected, so long as models used in this fashion are published. Second, discovery with models leaves clear data trails that can be re-inspected later (for example, Muldner and colleagues (2010) disagreed with Baker's (2007a) definition of gaming the system, and ran their alternate model on the same data set to compare the two models' predictions in detail). Third, some constructs that are difficult for humans to label by hand (such as latent student knowledge) can be modeled and then studied. Fourth, discovery with models enables analyses to occur at scales or for conditions that are otherwise infeasible. Collecting human labels of constructs such as the ones above can be very time-consuming; but once a model is developed, it can be quickly applied at a wide scale on additional data. For example, Baker & Gowda (2010) analyzed gaming the system and

off-task behaviors across an entire year of data from three schools, representing approximately a million data points.

The potential of discovery with models methods for studying learning and engagement in computer-based learning can be seen in the context of researching gaming the system (Baker, Corbett, & Koedinger, 2004), where a range of research that would be close to intractable to conduct through other methods has been possible for this construct. Models for gaming the system have been developed by multiple groups using both machine learning (Baker et al., 2004; Baker et al., 2008a; Baker, Mitrovic, & Mathews, 2010; Beal, Qu, & Lee, 2008; Walonoski & Heffernan, 2006) and knowledge engineering approaches (Aleven et al., 2006; Beck, 2005; Johns & Woolf, 2006; Muldner et al., 2011). These models have then been used to study how the prevalence of gaming changes during the year (Beck, 2005), what the relationships are between a range of motivational and attitudinal variables and gaming (Baker et al., 2008b; Beal, Qu, & Lee, 2008), the mechanisms through which gaming impacts learning (Cocea et al., 2009), whether gaming is more determined by the student or content (Baker, 2007a; Gong et al., 2010; Muldner et al., 2011), and the degree to which students in different settings game the system to different degrees (Baker & Gowda, 2010).

A Case Study on Discovery with Models: Carelessness and Student Motivation

In this paper, we demonstrate the potential of discovery with models for contributing to theory on learning through its use in researching carelessness. Despite accounts arguing for its prevalence and importance for almost a century (cf. Georges, 1929; McClure, 1929), carelessness has been lightly researched, in part due to the

complexity of operationalizing it. Generally, carelessness is described as giving the wrong answer despite having the needed skills for answering correctly. Better understanding the motivations which lead students to engage in careless behavior may lead to interventions that can reduce carelessness's frequency, or mitigate its negative effects.

Prior research has inferred carelessness based on an individual's performance on repeated assessments (Clements, 1982; Newman, 1977), observations and interviews (Georges, 1929), and self-report measures, or teacher's judgements, that treat carelessness as a stable learner characteristic (Hurlock & McDonald, 1934; Maydeu-Olivares & D'Zurilla, 1996; McClure, 1929; Rodríguez-Fornells & Maydeu-Olivares, 2000). In the behavioral assessment of carelessness used in (Newman, 1977; Clements, 1982), the student is administered the same item multiple times. This method is time-consuming and requires a special assessment outside of the regular learning task. It also may not be representative of carelessness occurring in other contexts: students might get bored of answering the same item, potentially inducing carelessness that would not have occurred otherwise.

Treating carelessness as a stable individual difference, besides reducing opportunities to understand the role of context in carelessness, is prone to demand effects (research subjects may perceive that there is an expectation of them to behave in a certain fashion) and self-presentation effects (answers given are influenced by research subject's desire to present himself or herself positively).

However, recent work in data mining and machine learning has made it possible to study carelessness within specific computer-based learning tasks, without modifying

the learning task, using automated carelessness detectors (cf. San Pedro et al., 2011) and discovery with models methods. We utilize this approach to study the relationships between carelessness and students' motivation, in particular their self-identified learning goals. In particular, we study the correlations between the prevalence of carelessness and student goals, and how student carelessness changed over time.

These issues are studied in the context of students engaging in scientific inquiry in a computer-based learning environment, Science Assistments (www.scienceassistments.org; Sao Pedro, Baker, Gobert, Montalvo, & Nakama, in press). In this context, carelessness is defined as failing to demonstrate a certain inquiry-related skill (specifically, control for variables strategy and hypothesis testing) despite knowing the skill (as measured by a Bayesian model of student knowledge). Carelessness is modeled by a machine-learned decision tree, which is used for discovering relationships between carelessness and student attributes (goals, beliefs and motivation).

Methods

Participants

Participants were 148 eighth grade students, 12-14 years old, from a public middle school in suburban Massachusetts. This school is majority White, with about 30% of the students from low-income families, and 30% on a free- or reduced-lunch program. The town where the school is located has a median income much higher than the U.S. median, and less than 2% of the families in the town are below poverty level (compared to about 10% in the US). School achievement in science is slightly lower than the state average, with 35% of the students categorized as proficient or higher on the state

standardized exam, the MCAS (state average = 40%). Students belonged to one of six class sections and had one of two science teachers. Students had no previous experience using microworlds with Science Assistments.

Materials and Design

Phase Change Microworld. The study was conducted in the context of students' scientific inquiry within a physical science "microworld", a virtual laboratory or simulation in which the student can run guided scientific experiments (cf. Sao Pedro et al., in press). In this study, we used a phase change microworld, where students explore how a particular substance, such as water, changes phases from solid to liquid to gas as it is heated. Each task in the microworld requires students to conduct experiments to determine if a particular independent variable (container size, heat level, substance amount, and cover status) affects various outcomes (melting point, boiling point, time to melt, and time to boil). For a given independent variable, students demonstrated proficiency by hypothesizing, collecting data, reasoning with tables and graphs, analyzing data, and communicating their findings. For this microworld, automated detectors of two key scientific inquiry skills – designing controlled experiment, and testing the stated hypothesis - were developed and validated (Sao Pedro et al., in press).

Motivation and goal orientation measures. Students completed the Patterns of Adaptive Learning Scales (PALS) survey (Midgley et al., 1996) in class as baseline data collected at the beginning of the school year, three months prior to using the phase change microworld. We analyze data from two scales capturing students' goal orientation, and their beliefs and attitudes towards learning. The first is the Personal

Achievement Goal Orientation (PALS1), including learning goal orientation, the goal of developing skill or learning (5 items), performance-approach goal orientation, the goal of demonstrating competence (5 items), and performance-avoid goal orientation, the goal of avoiding demonstrating incompetence (4 items). The second scale measures Academic-related Perceptions, Beliefs, and Strategies (PALS4), including academic efficacy (5 items), avoiding novelty (5 items), disruptive behavior (5 items), self-presentation of low achievement, the desire to prevent peers from knowing how well the student is performing (7 items), and skepticism about the relevance of school for future success (6 items). Each item has a 5-point Likert scale, and each of the above 8 sub-scales was calculated using the mean of each student's answers to the items for that sub-scale.

Procedure. After receiving a short introduction on using the tutor and on the activity, all students were engaged in the phase change learning activities over two class periods, about 70 minutes in total. During this time, Science Assistments logged all students' interactions within the learning environment. In the following sections, we show how we used these fine-grained interaction data to construct and validate machine-learned detectors of carelessness, based on existing detectors of systematic data collection behavior (Sao Pedro, Baker, Montalvo, Nakama, & Gobert, 2010; Sao Pedro et al., in press). Next, we analyze carelessness and its manifestation over consecutive activities, with regard to differences between students in motivation and goal orientation.

Carelessness Model Development

The process of building an automated detector of carelessness involves several steps. In the final step, a detector of carelessness is developed using machine learning. In

order to reach this step, we compute training labels for carelessness, developed based upon: a) Information on student correctness over time (e.g. whether the student succeeds or fails to demonstrate inquiry skills); and b) Information from an automated assessment of student knowledge, called Bayesian Knowledge Tracing (BKT) (cf. Corbett & Anderson, 1995). These labels of correctness/incorrectness are developed by aggregating students' actions in Science Assistments into "clips" (discussed next), which are labeled by machine-learned models of inquiry skill(s). These machine-learned models of inquiry skill(s) are derived from hand-coded assessments of correct or incorrect application of science inquiry skills. In the following sections, we discuss each of these steps.

Moving from fine-grained log files to coarse-grained clips. As students used Science Assistments, each individual's actions within the microworld were logged at a fine-grained level, with a focus on student actions while experimenting. Each action's type, current and previous values (where applicable – for instance, an independent variable's value before and after the action), and timestamp were recorded. In all, 27,257 relevant actions were logged, from 148 students. These data served as the basis for generating representations of student behavior that could be coded by humans in terms of inquiry skill, and then for generating machine-learned models of these inquiry skills.

Student behavior was segmented into a set of 1,503 clips, each of which consisted of all student experimentation between the time the student entered and left the experimentation phase of the microworld. (Students could have multiple clips for a specific activity if they chose to exit and re-enter experimentation, either to create new hypotheses or collect more data).

These clips were formatted as "text replays" (cf. Baker, Corbett, & Wagner, 2006), and labeled by two human coders based on whether a student demonstrated scientific inquiry skill during the scope of that clip (a binary variable). For example, one indication of skill in designing controlled experiments is the use of the Control of Variables Strategy (CVS), a method for creating experiments in which the value of a single variable is changed between consecutive steps (Chen & Klahr, 1999). However, other methods that generate unconfounded comparisons were also treated as evidence of skill in designing controlled experiments, for example modifying three variables and then changing two variables. The coders separately coded a second skill, whether the student tested their stated hypothesis (Collins & Stevens, 1991). The coders trained together on a set of clips, discussing their definitions of the construct based on specific examples and past literature (cf. Chen & Klahr, 1999; Collins & Stevens, 1991). They then coded clips separately to establish inter-rater reliability. The corpus of hand-coded clips contained exactly one randomly selected clip from each problem each student encountered, resulting in 571 clips; the two coders each coded the same sub-set of clips, and had acceptable inter-rater reliability (κ=.69 for designing controlled experiments, κ=1.00 for testing stated hypothesis).

The labels were then used as a basis for developing machine-learned models of these two inquiry skills (Sao Pedro et al., 2010; Sao Pedro et al., in press). The detectors produced were able to distinguish each of these behaviors 85% of the time (i.e., AUC=0.85 for each skill). Afterwards, a Bayesian Knowledge Tracing (BKT) model (Corbett & Anderson, 1995) was created to predict a student's latent knowledge of these

skills. The BKT models were able to predict future correctness on these skills over 70% over the time (AUC=0.74, 0.79).

Feature distillation. Each clip had a set of 73 features extracted for the machine-learning process (cf. Sao Pedro et al., 2010), including the numbers of different types of actions that occurred during the clip (including the number of complete and student-interrupted trials and the number of variable changes made while designing each experiment), the timing of each action (including the average time per variable change and the maximum time the student spent studying the simulation), and the probability that the student knew the skill involved with the relevant problem set before their first attempt on action N, P(Ln-1), calculated using Bayesian Knowledge Tracing (Corbett & Anderson, 1994).

Carelessness Detector. Our operational measure of carelessness is based on a detector of whether an incorrect answer was due to a lack of knowledge or not, based on features of the student response, previously termed "contextual slip" (cf. Baker, Corbett, & Aleven, 2008). This model was previously used to estimate carelessness in intelligent tutors for mathematics (Baker & Gowda, 2010; San Pedro et al., 2011). The term "slip" is used within multiple student modeling approaches, both in the Bayesian modeling literature and the psychometrics literature, to denote when the student makes an error despite knowing the requisite skills for correct performance (Corbett & Anderson, 1994; De La Torre & Douglas, 2004; Junker & Sijtsma, 2001; Morgan, 1979). This definition is essentially the same definition used for carelessness by Newman (1977). While slipping may not be a perfect indicator of carelessness (as a student could still obtain the correct answer despite being careless, if Maydeu-Olivares's definition is used rather than

Newman's), there are relatively few explanations for student errors on well-known skills, thereby making slipping a sound indicator of carelessness, if not a complete one. Slips can also imply a poor knowledge model (an issue studied by De La Torre & Douglas, 2004). We circumvent this possibility by studying slipping within a learning system that has a validated knowledge model to accurately predict future correctness 74% of the time (cf. Sao Pedro et al., in press), a level of correctness within the range of the current state of the art (cf. Pardos et al., 2011). Within our approach, training labels are computed by taking the probability that the student knew the skill prior to the attempt (computed using Bayesian Knowledge Tracing), and integrating this probability with data on future correctness on the next two problem solving attempts, using derivatives of Bayes' Theorem (the full mathematics for this process is outlined in Baker, Corbett, & Aleven, 2008). Next, machine learning is used to develop detectors that predict carelessness without using future data.

A detector predicting carelessness was developed using the REPTree (Reduced Error Pruning Tree) algorithm within the Weka extension package (Witten, Frank, & Hall, 2011) in RapidMiner 5.0 (Mierswa, Wurst, Klinkenberg, Scholz, & Euler, 2006). REPTree is a fast algorithm for constructing a decision tree for predicting the value of a numerical variable based on a set of other variables. Six-fold cross-validation was conducted at the student level (i.e., the detector is trained on five groups of students and tested on the sixth group), in order to assess whether the detector will remain accurate for new groups of students. An alternate algorithm, linear regression, was found to have systematic patterns in the model residuals for sub-sets of the data, which suggested that a non-linear model might be more effective. Indeed, REPTree outperformed linear

regression, achieving better cross-validated correlation (r=0.62 for REPTree, r=0.56 for linear regression), and Root Mean Squared Error (RMSE=0.16 for REPTree, RMSE=0.18 for linear regression).

The resulting REPTree had a size of 13 (including both interior nodes – decision points – and leaves – linear regression models) and a total depth of 6. Overall, the model used 3 variables:

- P(Ln-1), the probability the student knew the inquiry skill before the action
- Cumulated count of conducting two consecutive trials that vary in terms of only one variable
- Sum of timing for variable changes while constructing hypotheses

A prediction about carelessness was made for each clip, and then each student's predicted carelessness was computed by taking the average values over all of that student's clips. Overall, the mean carelessness probability across the studied clips (N=1282) was 0.08 (SD=0.14), which means that 8% of the errors across clips were detected as careless errors. When averaged across students, the mean carelessness was 0.12 (SD=0.16), which indicates that on average, 12% of students errors were detected as careless errors. This value represents considerably lower levels of carelessness than were seen in mathematics intelligent tutors or mathematics tests, in which carelessness levels were between 12%-50% (Baker & Gowda, 2010; Casey, 1979; Clarkson, 1983; Clements, 1980; Newman, 1977; San Pedro et al., 2011). Further exploration is needed to understand this difference.

Findings

The Relationship between Carelessness and Motivational Measures

In order to examine the relationships between careless and motivation, we calculated the correlations between carelessness and the 8 sub-scales of the PALS discussed above; the results are summarized in Table 1. Two of the sub-scales were marginally significantly correlated with carelessness:

a) Carelessness was marginally negatively correlated with *disruptive behavior*, r= -0.15, $F(1,128)=3.02$, $p=0.08$; i.e., more careless students tended to report less tendency to engage in disruptive behavior.

b) Carelessness was marginally negatively correlated with *self-presentation of low achievement*, r= -0.16, $F(1,128)=3.45$, $p=0.07$; i.e., those with higher scores on carelessness tended to have lower scores on *self-presentation of low achievement*.

Both disruptive behavior and self-presentation of low achievement have been previously shown to be related to low achievement (Ketter, 2006). Hence, one possible interpretation of the negative correlation between these constructs and carelessness is that these constructs lead to students making incorrect answers due to insufficient inquiry skill, instead of carelessness. We present an analysis of this possibility later in the paper.

As seen in Table 1, none of the three goal orientations assessed by PALS were statistically significantly correlated with carelessness ($r=0.08$ for learning goals, with $F(1,128)=0.68$, $p=0.36$; $r=0.04$ for performance-approach goals, with $F(1,128)=0.20$, $p=0.65$; and $r=0.001$ for performance-avoidance goals, with $F(1,128)=0.00$, $p=0.995$). However, it may be that a straightforward linear correlation does not capture the relationships between goal orientation and carelessness. Thus, to explore this possibility,

in the following section we examine whether there are more complex relationships between these constructs. Such complex relationships might be expressed when sub-sets of variables are considered (as opposed to single-variable-based calculations).

====== Insert Table 1 about here ======

Relationships between Carelessness and PALS-based Clusters

We can investigate the relationship between motivation and carelessness by looking at whether combinations of student motivations can predict carelessness. To analyze this, we conducted analyses to search for clusters of students who were similar in terms of responses on multiple motivational/goal measures, and studied whether there were differences in carelessness between these groups. This will be done using Cluster Analysis (Everitt, Landau, Leese, & Stahl, 2011), where data points (in this case - students) are assigned to data subsets (called clusters) based on inter-group similarities.

Based on a log-likelihood distance measure and the Bayesian Information Criterion (BIC; Raftery, 1995), using two clusters provided the optimal explanation of the data, with BIC = 705.6. However, using three clusters led to only minor degradation in BIC value, with BIC=711.9. The set of three clusters led to more interesting separations between aspects of the PALS (using two clusters essentially merged clusters 2 and 3); therefore we analyze the three-cluster model as shown in Table 2.

Cluster 1 (N=35) includes students who have high values for both *learning goal orientation* and *academic efficacy*, and can be referred to as the *learning goal orientation* cluster. Cluster 2 (N=66) includes students with (relatively) high values for both

performance-approach and *performance-avoid* goal orientations; these two types of goal orientation are often possessed by the same students (Bong, 2001; Darnon, Harackiewicz, Butera, Mugny, & Quiamzade, 2007; Elliot & Church, 1997; Middleton & Midgley, 1997; Ross, Shannon, Salisbury-Glennon, & Guarino, 2002). Though cluster 2 also has high values for learning goals, its values for performance goals are substantially higher than the other clusters. We will refer to cluster 2 as the *performance goal orientation* cluster. Finally, cluster 3 (N=20) includes students with high values for *avoiding novelty*, *disruptive behavior*, *self-presentation of low achievement*, and *skepticism about the relevance of school for future success*. Cluster 3's values for performance-avoidance goals are also almost as high as cluster 2's values, but the values for learning goals and performance-approach goals are substantially lower. As such, we will refer to this cluster as having a *lower tendency of either goal orientation*.

====== Insert Table 2 about here ======

The proportion of carelessness was marginally significantly different across the three clusters, $F(2,118)=3.05$, at $p=0.051$. The mean carelessness in cluster 3 (5%) was significantly lower than the mean carelessness in cluster 2 (12%), $t(82.3)=3.45$, $p<0.01$, or cluster 1 (16%), $t(38.9)=2.8$, $p<0.01$. No significant differences were found between clusters 1 and 2, $t(47.1)=1.0$, $p=0.32$. In all cases, the F of Levene's Test of Equality of Variances was significant at $p<0.05$, hence equal variances were not assumed.

Taken together, our results suggest that students with strong goal orientation towards learning (cluster 1) or performance (cluster 2) were on average twice as careless as those demonstrating a lower tendency to either type of goal orientation (cluster 3) - a surprising finding. One possible interpretation for this finding is that lower carelessness in students without learning or performance goals is due to these students having lower inquiry skills in general. By our operational definition, students completely lacking in a skill cannot demonstrate carelessness. Similarly, Clements (1982) found that carelessness was associated with better knowledge.

To address this possibility, we can compare students' inquiry skills over consecutive activities in each condition, measured as the mean values of $P(L_{n-1})$ – averaged over consecutive activities for each student – over the different clusters. $P(L_{n-1})$ was in fact significantly lower in cluster 3 (M = 12%, SD= 10%) than in cluster 1 (M = 24%, SD = 27%), $t(47.1)=2.4$, $p<0.05$, or cluster 2 (M = 22%, SD = 22%), $t(71.3)=2.7$, $p<0.01$. Hence, these results imply that the differences in carelessness between clusters may be due to the observed differences in student inquiry skills. Further evidence for this can be seen in the changes in carelessness over time in the three clusters. The sub-set of 106 students who completed the first three activities was taken, and carelessness was plotted over time. As shown in Figure 1, carelessness increases considerably over the course of the three activities for cluster 1 and cluster 2, but does not increase over time for cluster 3. Students in cluster 3 are not developing the degree of inquiry skill that makes careless errors feasible.

====== Insert Figure 1 about here ======

Discussion

In this paper, we have presented a case study in discovery with models: studying the motivations associated with carelessness in computer-based science inquiry. In discovery with models, a model is developed and applied to data, and then used as a component in other analyses, typically to discover aspects of the construct in the model. In this case, we developed and applied a machine-learned detector of student carelessness to log files, and correlated its outputs to motivational questionnaires, to discover the relationships between motivational measures and carelessness.

A key benefit of discovery with models is being able to study behavioral constructs in a non-disruptive fashion that is both scalable/longitudinal and fine-grained. In this specific case, using the automated detector allowed us to study carelessness in a more naturalistic and scalable fashion than was possible through interviews and repeated assessment of the same items (cf. Clements, 1982; Newman, 1977) and in a finer-grained fashion than was possible through questionnaire measures of carelessness (Maydeu-Olivares & D'Zurilla, 1996; Rodríguez-Fornells & Maydeu-Olivares, 2000). Carelessness is a difficult construct to label by hand, but can be inferred – as it is here – by the pattern of student correctness and the probability that student errors were not due to a lack of knowledge. In this paper, both the carelessness model and the student knowledge model used in analyses were explicitly validated in the data set where they were applied, an essential step for validly using a model in an analysis of this nature. Models can be applied more widely as well, if attention is paid to demonstrating generalizability for the type of transfer being conducted. Our carelessness model is based on the concept and

mathematics of Bayesian Knowledge Tracing, a widely-used algorithm for modeling student knowledge, which can be reconstructed in any learning environment in which students repeatedly interact with problems that have correct answers (or behaviors) that can be automatically detected. Hence, it is feasible to extend the work presented here to a variety of topics, domains, and populations.

Our results (the "discovery" in discovery with models), indicate that certain types of students are linked with carelessness, but simultaneously suggests a substantial role for student knowledge in carelessness. We find that student learning orientation can be described by three clusters: students characterized by high levels of learning goal orientation and academic efficacy, students characterized by high levels of both performance-approach and performance-avoid goals, and students lower in both types of goals. Students in the clusters characterized by learning or performance goals have (on average) double the probability of carelessness as students in the third cluster. One potential interpretation is that students with higher mastery or performance goals succeed in learning and correspondingly become more confident, as suggested by Clements (1982), and that this confidence leads to carelessness, despite their identified goal orientations. Our results suggest that students in these two clusters achieve higher learning, particularly over time, and that this factor may drive the differences in carelessness. It is also worth noting that academic efficacy, also measured by PALS (Midgley et al., 1996), is higher in these two clusters.

Surprisingly, carelessness was substantially less common in our dataset than in previous research (cf. Baker & Gowda, 2010; Clements, 1982; San Pedro et al., 2011). It may be that students solving problems in mathematics – the domain used by Baker

(Baker & Gowda, 2010; San Pedro et al., 2011) and by Clements (1982) – tend to be more careless than students in science inquiry, due to differences between these two domains. By its very nature, scientific inquiry is exploratory and involves active self-direction in searching for knowledge (Anderson, 2002; Haury & OH, 1993; Novak, 1964), potentially reducing carelessness. Alternatively, some aspect of the science microworld or its novelty may have reduced the amount of carelessness exhibited.

The relationships found here between goal orientation and carelessness worth comparing to previous research on other disengaged behaviors, much of it also conducted with a discovery with models approach. Neither gaming the system nor off-task behaviors were found to be correlated with performance and learning goals (Baker et al., 2007b; 2008b). Hence, our results provide somewhat of a contrast to these previous results. One difference between carelessness and these behaviors may be the situations in which carelessness occurs: While gaming often occurs when a student has not yet learned a skill (Baker et al., 2004) and off-task behavior is distributed throughout the learning experience, careless errors occur after a skill has been learned. Similarly, other measures of disengagement appear to have different relationships to goal orientation than seen here for carelessness. Luo and colleagues (2011) found that students with performance-approach goals report paying more attention to work in class and homework than students with neither learning goals nor performance-approach goals (e.g., the students corresponding to the third cluster). These findings are seemingly inconsistent with the findings in this paper; one possible explanation is that the different results stem from using a very different measurement approach (self-report questionnaires), or the difference in domain. Overall, these results demonstrate that carelessness and learning

orientation have surprising relationships, which merit further investigation. Understanding the factors that lead successful students to become careless may support the design of interventions to assist these students in maintaining high performance even when they have learned the material.

More broadly, this paper demonstrates the potential of discovery with models analyses for conducting analyses of learning behaviors that are difficult to study with more traditional methods. In the current paper, the learning behavior examined is carelessness, the model used to study it is a machine-learned decision tree, and our "discovery" is evidence on the relationships between a student's carelessness and his or her goals, beliefs and attitudes. By using discovery with models in combination with other methods, we may be able to shed light on questions that have thus far been intractable. As such, we view discovery with models as a key learning analytics method for 21st-century learning science.

Acknowledgements

This research was funded by NSF #DRL-1008649, "Empirical Research: Emerging Research: Using Automated Detectors to Examine the Relationships Between Learner Attributes and Behaviors During Inquiry in Science Microworlds" awarded to Janice Gobert and Ryan Baker. Any opinions are those of the authors and do not necessarily reflect those of the funding agency. We would like to thank Matthew Bachmann for his help in data coding.

References

Aleven, V., Mclaren, B., Roll, I., & Koedinger, K. (2006). Toward meta-cognitive tutoring: A model of help seeking with a Cognitive Tutor. *International Journal of Artificial Intelligence in Education, 16*(2), 101-128.

Anderson, R. D. (2002). Reforming science teaching: What research says about inquiry. *Journal of Science Teacher Education, 13*(1), 1-12.

Arnold, K. E. (2010). Applying academic analytics. *Educause Quarterly, 33*(1).

Arroyo, I., Woolf, B. P., Cooper, D., Burleson, W., & Muldner, K. (2011). The impact of animated pedagogical agents on girls' and boys' emotions, attitudes, behaviors, and learning. *Proceedings of the 11th IEEE Conference on Advanced Learning Technologies,* 506-510.

Baker, R. S. J. d. (2007a). Is gaming the system state-or-trait? Educational data mining through the multi-contextual application of a validated behavioral model. *Complete On-Line Proceedings of the Workshop on Data Mining for User Modeling at the 11th International Conference on User Modeling 2007, 76-80.*

Baker, R. S. J. d. (2007b). Modeling and understanding students' off-task behavior in intelligent tutoring systems. *Proceedings of ACM CHI 2007: Computer-Human Interaction, 1059-1068.*

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. *Proceedings of the 9th International Conference on Intelligent Tutoring Systems, 406-415.*

Baker, R. S. J. d., 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.*

Baker, R. S. J. d., Corbett, A. T., Roll, I., & Koedinger, K. R. (2008a). Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction, 18*(3), 287-314.

Baker, R. S. J. d., Corbett, A. T., & Wagner, A. Z. (2006). Human classification of low-fidelity replays of student actions. *Proceedings of the Educational Data Mining Workshop at the 8th International Conference on Intelligent Tutoring Systems, 29-36.*

Baker, R. S. J. d., & Gowda, S. M. (2010). An analysis of the differences in the frequency of students' disengagement in urban, rural, and suburban high schools. *Proceedings of the 3rd International Conference on Educational Data Mining, 11-20.*

Baker, R.S.J.d., Mitrovic, A., Mathews, M. (2010). Detecting gaming the system in constraint-based tutors. *Proceedings of the 18th Annual Conference on User Modeling, Adaptation, and Personalization*, 267-278.

Baker, R. S. J. d., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., & Koedinger, K. (2008b). Why students engage in "gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research, 19*(2), 185.

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*(1), 3-17.

Beal, C.R., Qu, L., Lee, H. (2008) Mathematics motivation and achievement as predictors of high school students' guessing and help-seeking with instructional software. *Journal of Computer Assisted Learning, 24*(6), 507-514.

Beck, J. E. (2005). Engagement tracing: Using response times to model student disengagement. *Proceedings of the International Conference on Artificial Intelligence and Education (AIED2005), 88-95.*

Bong, M. (2001). Between-and within-domain relations of academic motivation among middle and high school students: Self-efficacy, task value, and achievement goals. *Journal of educational psychology, 93*(1), 23.

Casey, D. (1979). An analysis of errors made by junior secondary pupils on written mathematical tasks'. *Unpublished Master of Education thesis, Monash University*.

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development, 70*(5), 1098-1120.

Clarkson, P. (1983). Types of errors made by Papua New Guinean students. *Educational Studies in Mathematics, 14*(4), 355-367.

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

Clements, M. (1982). Careless errors made by sixth-grade children on written mathematical tasks. *Journal for Research in Mathematics Education, 13*(2), 136-144.

Cocea, M., Hershkovitz, A., & Baker, R. S. J. (2009). The impact of off-task and gaming behaviors on learning: immediate or aggregate? *Proceedings of the 14th International Conference on Artificial Intelligence in Education, 507-514.*

Collins, A., & Stevens, A. L. (1991). A cognitive theory of inquiry teaching. *In P. goodyear (Ed.), Teaching Knowledge and Intelligent Tutoring (pp. 203-230).* Norwood, NJ: Ablex.

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-adapted Interaction, 4*(4), 253-278.

Darnon, C., Harackiewicz, J. M., Butera, F., Mugny, G., & Quiamzade, A. (2007). Performance-approach and performance-avoidance goals: When uncertainty makes a difference. *Personality and Social Psychology Bulletin, 33*(6), 813-827.

Dawson, S., Macfadyen, L., Lockyer, L., & Mazzochi-Jones, D. (2011). Using social network metrics to assess the effectiveness of broad-based admission practices. *Australasian Journal of Educational Technology, 27*(1), 16-27.

De La Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333-353.

Elliot, A. J., & Church, M. A. (1997). A hierarchical model of approach and avoidance achievement motivation. *Journal of Personality and Social Psychology, 72*, 218-232.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis (5th Ed.).* West Sussex, UK: Wiley.

Georges, J. (1929). The nature of difficulties encountered in reading mathematics. *The School Review, 37*(3), 217-226.

Gong, Y., Beck, J., Heffernan, N. T., & Forbes-Summers, E. (2010). The impact of gaming (?) on learning at the fine-grained level. *Proceedings of the 10th International Conference on Intelligent Tutoring Systems, 194-203.*

Haury, D. L., & OH, E. C. (1993). Teaching Science through Inquiry. ERIC/CSMEE Digest. *Columbus, OH: ERIC Clearinghouse for Science, Mathematics and Environmental Education. Retrieved September, 25*, 2008.

Hurlock, E., & McDonald, L. (1934). Undesirable behavior traits in junior high school students. *Child Development, 5*(3), 278-290.

Jeong, H., & Biswas, G. (2008). Mining student behavior models in learning-by-teaching environments. *Proceedings of the 1st International Conference on Educational Data Mining*, 127-136.

Johns, J., Woolf, B.P. (2006) A dynamic mixture model to detect student motivation and proficiency. *Proceedings of the National Conference on Artificial Intelligence,* 163-168.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 258-272.

Ketter, L. C. (2006). High-stakes Testing, Achievement-goal Structures, Academic-related Perceptions, Beliege, Strategies, and School Belonging Among Selected Eighth-grade Students in a Northwest Florida School District. *Unpublished Doctor of Education Dissertation, The University of West Florida.*

Kinnebrew, J. S., Biswas, G., & Sulcer, B. (2010). Modeling and measuring self-regulated learning in teachable agent environments. *Journal of e-Learning and Knowledge Society, 7*(2), 19-35.

Luo, W., Paris, S. G., Hogan, D., & Luo, Z. (2011). Do performance goals promote learning? A pattern analysis of Singapore students' achievement goals. *Contemporary Educational Psychology, 36*(2), 165-176.

Maydeu-Olivares, A., & D'Zurilla, T. J. (1996). A factor-analytic study of the Social Problem-Solving Inventory: An integration of theory and data. *Cognitive Therapy and Research, 20*(2), 115-133.

McClure, W. (1929). Characteristics of problem children based on judgments of teachers. *Journal of Juvenile Research, 13*, 124-140.

Middleton, M. J., & Midgley, C. (1997). Avoiding the demonstration of lack of ability: An underexplored aspect of goal theory. *Journal of Educational Psychology, 89*(4), 710-718.

Midgley, C., Maehr, M. L., Hicks, L., Roeser, R., Urdan, T., Anderman, E. M., & Kaplan, A. (1996). *The Patterns of Adaptive Learning Survey (PALS)*. Ann Arbor, MI: University of Michigan Press.

Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., & Euler, T. (2006). YALE: Rapid prototyping for complex data mining tasks. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 935–940.*

Militello, M., Heffernan, N. (2009). Which one is "just right"? What every educator should know about formative assessment systems. Kansas City, MO: *National Council of Professors of Educational Administration.*

Morgan, G. (1979). *A Criterion-referenced Measurement Model with Corrections for Guessing and Carelesness*: Australian Council for Educational Research.

Muldner, K., Burleson, W., Van de Sande, B., & VanLehn, K. (2011). An analysis of students' gaming behaviors in an intelligent tutoring system: predictors and impacts. *User Modeling and User-Adapted Interaction, 21*(1), 99-135.

Newman, M. A. (1977). An analysis of sixth-grade pupils' errors on written mathematical tasks. *Victorian Institute for Educational Research Bulletin, 39*, 31-43.

Novak, A. (1964). Scientific inquiry. *Bioscience*, 25-28.

Pardos, Z.A., Baker, R.S.J.d., Gowda, S.M., & Heffernan, N.T. (2011). The sum is greater than the parts: Ensembling models of student knowledge in educational software. *SIGKDD Explorations, 13*(2), 37-44.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111-164.

Rodríguez-Fornells, A., & Maydeu-Olivares, A. (2000). Impulsive/careless problem solving style as predictor of subsequent academic achievement. *Personality and Individual Differences, 28*(639), 639-645.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications, 33*(1), 135-146.

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 40*(6), 601-618.

Ross, M. E., Shannon, D. M., Salisbury-Glennon, J. D., & Guarino, A. (2002). The patterns of adaptive learning survey: A comparison across grade levels. *Educational and Psychological measurement, 62*(3), 483-497.

San Pedro, M., Baker, R. S. J. d., & Rodrigo, M. (2011). Detecting carelessness through contextual estimation of slip probabilities among students using an intelligent tutor for mathematics. *Proceedings of 15th International Conference on Artificial Intelligence in Education*, 304-311.

Sao Pedro, M. A., Baker, R. S. J. d., Montalvo, O., Nakama, A., & Gobert, J. D. (2010). Using text replay tagging to produce detectors of systematic experimentation behavior patterns. *Proceedings of the 3rd International Conference on Educational Data Mining, 181-190.*

Sao Pedro, M. A., Baker, R. S. J. d., Gobert, J., Montalvo, O., & Nakama, A. (in press). Leveraging machine-learned detectors of systematic inquiry behavior to estimate and predict transfer of inquiry skill. *User Modeling and User-Adapted Interaction*.

Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *Educause Review, 46*(5), 30-40.

Walonoski, J., & Heffernan, N. (2006). Detection and analysis of off-task gaming behavior in intelligent tutoring systems. *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, 382-391.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical machine learning tools and techniques (3rd Ed.)*. Burlington, MA: Morgan Kaufmann.

Table 1

Correlation coefficients between carelessness and PALS measures (N=130)

PALS Measure	**r**
Learning goal orientation	0.08
Performance-approach goal orientation	0.04
Performance-avoid goal orientation	0.001
Academic efficacy	0.14
Avoiding novelty	-0.13
Disruptive behavior	*-0.15*[†]
Self-presentation of low achievement	*-0.16*[††]
Skepticism about school relevance for future success	-0.12

[†] p=0.08, [††] p=0.07

Table 2

Center values (means) of each construct for the clusters. Cells in gray represent the cluster with the highest value for the current variable.

Variable	**Mean (SD)**			**Difference Between Groups F(2,118)**	**η^2 (between-group effect size)**
	Cluster 1	**Cluster 2**	**Cluster 3**		
Learning goal orientation	4.66 (0.40)	4.38 (0.64)	2.07 (0.87)	46.49**	0.44
Performance-approach goal orientation	1.69 (0.57)	3.20 (1.04)	2.40 (0.82)	36.14**	0.38
Performance-avoid goal orientation	1.86 (0.72)	3.78 (0.67)	3.62 (0.68)	92.55**	0.61
Academic efficacy	4.41 (0.49)	4.22 (0.55)	3.65 (1.06)	13.13**	0.18
Avoiding novelty	1.96 (0.60)	2.58 (1.00)	3.02 (1.21)	21.76**	0.27
Disruptive behavior	1.54 (0.68)	1.61 (0.68)	2.07 (1.01)	27.57**	0.32
Self-presentation of low achievement	1.33 (0.31)	1.59 (0.60)	3.43 (1.00)	8.93**	0.13
Skepticism about the relevance of school for future success	1.57 (0.49)	1.92 (0.82)	2.07 (0.87)	39.17**	0.40
N	**35**	**66**	**20**		
Carelessness	**0.16 (0.22)**	**0.12 (0.13)**	**0.05 (0.05)**		

** $p<0.01$

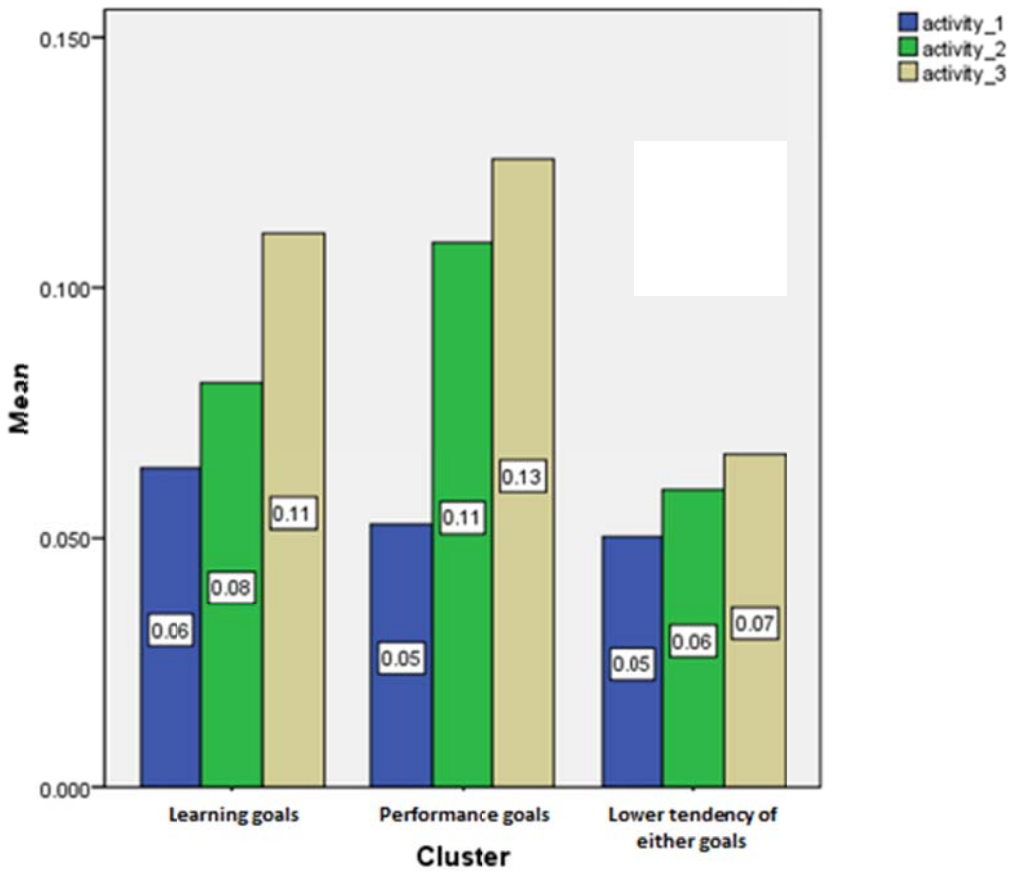

Figure 1. Students' mean carelessness values over consecutive activities by clusters