Detecting Diligence with Online Behaviors on Intelligent Tutoring Systems

Steven Dang
HCI Institute
Carnegie Mellon University
Pittsburgh, PA, USA
stevenda@cs.cmu.edu

Michael Yudelson
HCI Institute
Carnegie Mellon University
Pittsburgh, PA, USA
yudelson@cs.cmu.edu

Kenneth R. Koedinger
HCI Institute
Carnegie Mellon University
Pittsburgh, PA, USA
koedinger@cmu.edu

ABSTRACT

The current study introduces a model for measuring student diligence using online behaviors during intelligent tutoring system use. This model is validated using a full academic year dataset to test its predictive validity against long-term academic outcomes including end-of-year grades and total work completed by the end of the year. The model is additionally validated for robustness to time-sample length as well as data sampling frequency. While the model is shown to be predictive and robust to time-sample length, the results are inconclusive for robustness in data sampling frequency. Implications for research on interventions, and understanding the influence of self-control, motivation, metacognition, and cognition are discussed.

Author Keywords

Self-Control; Self-Regulated Learning; Intelligent Tutoring Systems; Measurement; Noncognitive factors; Learning Analytics; Diligence; Motivation; Online Behaviors

INTRODUCTION

The oft-cited 10,000 hour rule, popularized by Malcolm Gladwell as the amount of time required to build expertise, does not completely describe an amateur's pathway to mastery [23]. Not just any practice will lead to expertise; practice that is at the edge of students' abilities will be most effective at improving abilities. This type of practice is typically referred to as "deliberate practice" [14], and because it demands the student to perform at the limits of his or her existing abilities, such practice can tax the student's mental and physical resources. Thus deliberate practice requires students to constantly regulate their learning and exercise self-control to remain focused. As learning shifts increasingly towards digital environments, students will be tempted by more distractions and it is important that they resist and remain diligent while learning.

In the past decade, there has been mounting evidence that self-control influences long-term academic outcomes[29]. In 2013, the US Department of Education released a report summarizing the evidence supporting the role of self-control and similar non-cognitive factors in academic performance. This new area of interest has led educators to push for interventions that promote greater self-control during learning. As these interventions proliferate, so does research to better understand their efficacy and interactions with other factors including motivation, metacognition, and cognition. Research in this field relies on survey-based measures of self-control [29]. Furthermore, with the push to include assessment as part of state standards, there is a need for a more robust measure [10]. This has created a demand for a validated behavioral measure to complement existing measures.

Digital courses present an opportunity to explore whether behavioral measures of self-control can be computed from the fine-grained, high-volume data generated by student actions in more interactive courses. If so, there is a great potential benefit relative to survey or specially-constructed behavioral assessments. Such a model could unobtrusively detect student levels of self-control as a natural consequence of course interaction. Key questions for such an exploration are what models can effectively convert raw interaction data into self-control measures and at what scale must data be collected, particularly in terms of the observations per student, for such measures to be reliable and have predictive validity.

In this study, we operationalize student diligence, a facet of self-control, and introduce a model for measuring student diligence using behaviors logged while learning with an intelligent tutoring system. We validate this measure using a year-long large-scale dataset (2.5 million observations) that has a modest scale in terms of students (108), but a large scale in terms of observations per student (about 15,000). The long time frame facilitates analysis of the measure's robustness variance in the time-sample length and volume of per student data. We also assess the convergent and divergent validity of the diligence measure with other self-regulated learning and self-control constructs assessed through associated surveys.

L@S 2017, April 20-21, 2017, Cambridge, MA, USA
 ISBN 978-1-4503-4450-0/17/04…$15.00
DOI: http://dx.doi.org/10.1145/3051457.3051470

BACKGROUND

What is Diligence

Diligence has been defined as working assiduously on academic tasks which are beneficial in the long-run but tedious in the moment, especially in comparison to more enjoyable, less effortful diversions[16]. Thus, diligence is the domain-specific ability to maintain a high degree of focus on a given task within that domain. This highlights two important relations to the higher-level construct of self-control, domain-specificity and trait-like stability. Self-control is generally considered trait-like and is the broader ability of an individual to regulate emotions, behaviors, and thoughts especially under the temptation of desirable alternatives [30]. This trait-like quality is driven by the control facet of the "Big Five" personality trait, conscientiousness [22]. Therefore diligence should be stable across contexts in aggregate. However, while self-control should be relatively stable aggregated across contexts and domains, like many traits, it can have relatively low correlations between domains(r=.20-.30) [31]. Thus math diligence may vary greatly from athletic diligence, and it is important to measure the domain-specific self-control displayed to accurately capture its influence. This definition of diligence follows directly from the social and developmental psychology literatures which also use the terms willpower [24] and ego-resiliency [17] to refer to self-control. Alternatively, the cognitive and educational psychology communities have similar constructs that have been enumerated specifically for learning contexts. We describe some of these constructs next. We do so both because they give a sense of the rich landscape of related (hypothesized) psychological constructs and because the data we analyze includes survey measures corresponding not only to diligence but also to these other constructs.

Relation to Executive Function

Executive function is a cognitive function heavily implicated in self-control [5]. Similar to personality measures such as conscientiousness, executive function is a relatively constant cognitive resource that consists of three components: inhibitory control, working memory, and set shifting. All of these components are recruited collectively in diligent behaviors. However, associations between low-level executive function measures and self-control in real-world tasks are small [9].

Relations to Self-Regulated Learning

Self-regulated learning (SRL) is a framework from the education community that subsumes and integrates a wide range of beliefs, skills, and strategies that impact learning and originate from the self [33]. This framework views self-regulation as a set of motivational, metacognitive, and behavioral constructs that drive a plan-act-reflect cycle. Each construct has its own specific moderating relationship with self-control and diligence.

Self-Efficacy

Self-efficacy is defined as the belief in one's ability to perform at a given level on a range of tasks [4]. Self-Efficacy moderates self-control and thus ratings of self-efficacy should be correlated to measures of diligence.

Achievement Goals

Achievement goals are a cognitive knowledge construct defined by a 2x2 matrix where one axis is Performance versus Mastery and the other is Approach versus Avoidance [27]. Performance goals are those that define accomplishment relative to peer-derived standards while mastery goals are ones that are defined relative to personal standards and prior ability and knowledge. Approach orientation implies an individual is seeking attainment of those goals while avoidance orientation describes individuals more concerned with avoiding failure rather than goal attainment. Thus mastery approach describes individuals who work towards attaining greater knowledge or ability and performance avoidance describes individuals who work to avoid having lower grades than their peers [13]. As a cognitive knowledge construct, these are activated by the task context to moderate self-control. While mastery orientation should tend to focus executive function on task specific information, performance goals will have more variable influences on self-control depending on the dynamics of performance and the resulting behaviors and strategies employed [2].

Theory of Intelligence

Theory of Intelligence describes a mindset related to the nature of human intelligence [12]. Fixed mindset describes an individual's belief that intelligence and thus academic accomplishment is a fixed and predetermined trait (i.e. some people are smart others are not). Growth mindset describes the belief that intelligence can be developed. While it appears that mindsets should have an influence on self-control through ego-depletion [18], it turns out that mindsets are uncorrelated with conscientiousness [8] and thus are likely uncorrelated with diligence.

Effort Regulation

Effort Regulation is one scale from the MSLQ Self-regulated learning inventory [26]. It is defined as a students' ability to control their effort and attention in the presence of distractors. This construct is analogous to diligence as defined, and thus should be highly correlated.

Existing Measures

Each executive function has an associated behavioral task that has been validated such as the star-counting task measuring working memory. However, the predictive validity of such measures is low due to the weak association between these low level cognitive measures and more complex real-world tasks involving self-control [25].

SRL researchers have made a number of notable strides towards online measurements instruments. [32] created a note-taking, collaborative study-aid tool. This system found student judgments of their learning process to not reliably match assessments of online behaviors, thus raising questions of measured construct validity. [1] was able to identify help-seeking strategies as informed by SRL theories using log data collected from an intelligent tutoring system. [15] found conditional but no direct links in looking for online measures of achievement goals based on online hint-seeking and glossary-use behaviors.

For self-control measures, there are currently no online behavioral measures, but [16] introduce a math based behavioral task that served as inspiration for our investigation. In [16], the authors introduce the Academic Diligence Task (ADT), a math-based task that is targeted for high-school aged students and older. The ADT attempts to measure diligence by monitoring how long students engage in a tedious but beneficial math task versus a more immediately rewarding alternative, playing video games and watching videos. They are told "try to solve as many problems as quickly and accurately as you can" and "you are doing this activity because it can make you smarter" to create the expectation that they should do the math task and that it is good for them. More specifically, students are asked to solve single-digit subtraction problems for 4 five-minute windows. The computer interface is split between a math problem interface and video-watching/game-playing interface. During this task, the total time spent solving math problems as opposed to watching videos or playing games is logged. Also the total number of problems solved is logged. These two measures were correlated with self-control and conscientiousness, but not with other big five personality traits. They were also predictive of long-term outcomes including end-of-year grade, graduation, and 4-year college admission.

Adapting the Diligence Model for a Cognitive Tutor

The ADT utilized a low skill task in order to tax the mental facilities associated with self-control such that more diligent students would tend to stay on task more often, while less diligent students would tend to stray from the task. Thus more time on task and more problems solved translated to greater diligence.

The model proposed by the ADT would be as follows in *(1)*:

$$\mathbf{Y}_{dil} = \beta_0 \mathbf{X}_{tot} + \beta_1 \mathbf{X}_{prod} + \varepsilon \qquad (1)$$

Where $\mathbf{Y}_{dil}$ is the measured diligence. $\mathbf{X}_{tot}$ is the total time on task. $\mathbf{X}_{prod}$ is the total number of correct problems completed. $\boldsymbol{\varepsilon}$ is a Gaussian random error term.

By design, this task is able to differentiate the diligence of students who are very fluent in simple arithmetic, however, it is less likely to be able to differentiate students in the 1st grade who are only just learning how to subtract single digit numbers. Thus a more general model of student diligence would be valuable to assess a wider range of students.

Students are increasingly using highly interactive online course materials, such as intelligent tutoring systems, and logged student interactions with these systems are a rich source of student behavior during learning. The availability of such data provides an opportunity to explore whether these observations of naturalistic student learning behaviors can be utilized for the assessment of diligence. There are several challenges to using a cognitive tutor as a diligence assessment in place of the ADT. The first challenge is that cognitive tutors are designed to be adaptive to student's knowledge, moving on to the new material upon reaching mastery [28]. Thus students solving the same number of problems may actually have learned different amounts. Similarly, errors during learning will adjust the knowledge model and lead towards increased practice on a given problem type. Thus the raw number of problems completed, as proposed by the ADT, is not as directly comparable across students.

In contrast to the ADT, another challenge with using intelligent tutoring systems is that students are solving problems using a variety of learning processes including deliberate sense-making, inductive learning, and fluency-building cognitive processes [20]. In the ADT, the simple nature of the problem reduced the cognitive load to a fluency-building task, where time per problem solved should be nearly constant throughout the task. While working on cognitive tutors, students may pause for productive reasons such as reflection or sense-making [1], as well as for unproductive reasons such as socializing [3]. Thus interpreting time-on-task is trickier than it is in the ADT.

As with any cognitive task, greater prior-knowledge is going to enable superior task performance. Thus, this is also likely a factor that will have to be taken into account. Students with greater math ability will tend to solve more problems in the same amount of time as their peers with less ability.

Taking these factors into account, the following model is proposed as shown in *(2)*.

$$\mathbf{Y}_{dil} = \beta_0 \mathbf{X}_{tot} + \beta_1 \mathbf{X}_{work} + \beta_2 \mathbf{X}_{prior} + \varepsilon \qquad (2)$$

Where $\mathbf{Y}_{dil}$ is the measured diligence. $\mathbf{X}_{tot}$ is the total time in the system as a sum of the duration of all steps in the sampled time period. $\mathbf{X}_{prod}$ is the total number of correct steps completed in the sampled time. $\mathbf{X}_{work}$ is the average work rate as computed by $\mathbf{X}_{prod}$ / $\mathbf{X}_{tot}$. $\mathbf{X}_{prior}$ is the prior knowledge of the student, which in this work is equivalent to the grade from the previous year's math course. $\boldsymbol{\varepsilon}$ is a Gaussian random error term.

Fitting the model

For each model (1) and (2), diligence is assumed to be some linear combination of the measured behaviors from the intelligent tutoring system. In order to learn the context specific coefficients of these parameters, the normalized number of curriculum units mastered at the end of the year

by each student is used in place of Y_{dil}. Thus equations (3) and (4) are utilized to learn the values of β_0, β_1, and β_2 for (1) and (2) respectively. This defines the model as the components of the online behaviors along the student learning latent subspace.

$$Y_{out} = \beta_0 X_{tot} + \beta_1 X_{prod} + \varepsilon \quad (3)$$

$$Y_{out} = \beta_0 X_{tot} + \beta_1 X_{work} + \beta_2 X_{prior} + \varepsilon \quad (4)$$

Where $\mathbf{Y}_{out}$ is the number of curriculum units completed by the student.

The justification for this model is best understood with a few examples.

Varying Time-on-task

Comparing two students with the similar prior knowledge and who have been solving problems equally fast, the student who spends more time solving problems instead of quitting the application early, delaying getting started at the beginning of class, or taking more bathroom breaks, is the more diligent student.

Varying Work-Rate

Comparing two students with the similar prior knowledge and who have been solving problems for the same amount of time, the student who is solving more problems is likely doing so because they are focusing more and learning more per problem as a result. This makes the faster working student more diligent due to their increased exercising of self-control to focus on the task at hand.

Varying Prior-Knowledge

Comparing one student with less prior knowledge to one with more prior knowledge, the social pressure of the class context may encourage students to reach certain milestones. Students with more knowledge may feel less pressure to work as quickly, but when they are working just as fast and just as long as less knowledgeable students, they are demonstrating greater diligence.

THIS STUDY

In this study, we look to validate the proposed model as having superior model fit over the ADT for data from adaptive learning environments. We then characterize the predictive validity of the model for end-of-year grade and amount of material completed by the end of the year. We characterize the robustness of the model to data sampling from varying time-grain sizes. We look at convergent and discriminant validity with other motivation and metacognition constructs collected through surveys. Finally we finish with an analysis looking at the predictive validity of varying time-grains with sparser samples to characterize a lower bound on data required to support this model.

DATASET

This dataset [6] includes over 2.5M transactions from 108 students middle school students in pre-algebra class using a Carnegie Learning's Cognitive Tutor on a regular basis (two class-periods/week for the entire year). The data was collected as part of a different study [7], but we have utilized here because it includes a long time-window as well as motivational and metacognitive survey measures. The students are all from a single middle class suburban school in a mid-Atlantic state. The dataset includes 87 seventh graders and 21 eighth graders. There are equal numbers of male and female students, and the population is predominantly Caucasian, with 104 Caucasian and 4 non-Caucasian students.

Carnegie Learning's Cognitive Tutor

While the model introduced is designed for more general online behavioral assessment, this study leverages Carnegie Learning's Cognitive Tutor (CT) dataset for the aforementioned reasons. The CT utilized in this dataset is an Intelligent Tutoring System for Pre-Algebra that is deployed across thousands of middle schools across the United States. The CT leverages computational cognitive models to provide adaptive problem selection and hint support and correctness feedback to the students. Problems are broken down into a multi-step process, which allows the system to identify independent skills and trace skill improvement over a fine-grained skill model of the domain. The system logs all interactions with the system including problem attempts, hint requests, response accuracy, and problem step time. In this study, transactions for all students over the course of an entire academic year are utilized.

Collected Metadata Measures

In addition to online behavior logs from the CT, each student's course grades for the previous year, each academic quarter, and the end-of-year course grade are reported alongside several surveys of motivation and metacognition completed at the beginning of the academic year before any course content was completed.

Self Efficacy

Self-efficacy was measured using a 5-question scale with a 5-point Likert rating assessing student's self-efficacy with respect to their performance in the math class.

Achievement Goal Orientation

Achievement goals for Mastery Approach, Performance Approach, and Performance Avoidance were assessed using the corresponding 9 questions from the AGQ-R 12-question scale [13] with a 5-point Likert rating.

Theory of Intelligence

Theory of Intelligence was assessed using a 6-question scale from [12] reported using a 5-point Likert rating.

Effort Regulation

Effort Regulation is measured using the 4-question scale from the MSLQ [26] using a 5-point Likert rating.

RESULTS

Comparing Model Fit

The proposed model was compared to the ADT model in order to determine which model had better fit to the data in the intelligent tutoring context. Linear fixed effect models

Model	AIC	BIC
(1) ADT Model	300.00	308.03
(2) Proposed Model	298.13	306.15

Table 1. Model fit for data from full academic year.

were constructed according to equations *(1)* and *(2)* with each proposed variable as fixed effects, intercepts were removed as insignificant extrapolations of the data, and number of units completed over the entire year was set as the dependent variable.

Both the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) were calculated for each model. The AIC and BIC values for each model are shown in Table 1. In both cases, the proposed model is found to have better fit to the data, and thus all analysis moving forward was conducted using the proposed model.

Predictive Validity

We then tested the predictive validity of our diligence measure for both curriculum units completed in a year and end-of-year course grade using ordinary least-squares regression. In both models, gender, ethnicity, free or reduced lunch, interest in math, and previous math achievement are controlled for. All variables are normalized in order to facilitate interpretation of coefficients.

The results of the regressions are shown in Table 2. The diligence measure was predictive of both Final Grade (R^2=0.53) and Units Completed (R^2=0.62).

In order to gain insight into the nature of the models' predictions, the actual outcome measures and estimated outcome measures were divided into quintiles and the type and size of errors made by the model were analyzed. Tables 3 and 4 show the accuracy and error rate of the model for End-of-year Grade and Units Completed respectively. As expected given the models' R^2 values, there is a strong diagonal to both matrices implying both high accuracy and small magnitude errors at each level. One notable feature is that the models more accurately predict the correct class at the bottom and top quintiles(66% on average) as opposed to the middle three quintiles (35% on average).

Parameter	Final Grade β(p-value)	Units Completed β(p-value)
Gender	0.06(.58)	0.06(.55)
F&R Lunch	-0.17(.25)	-0.03(.80)
Ethnicity	0.28(.41)	0.26(.40)
Math Interest	0.12(.073)	0.15(.017)*
Prior Grade	0.20(<0.01)**	0.07(.217)
Diligence	1.89(<0.001)***	1.64(<0.001)***

Table 2. Regression model using full academic year data.

End-of-Year Grade	1st 20%	2nd 40%	3rd 60%	4th 80%	5th 100%
Correct Pos.	55%	24%	38%	33%	68%
Type I	12%	17%	14%	19%	8%
Type II	45%	76%	62%	67%	77%
Correct Neg.	88%	83%	86%	81%	80%

Table 3. Model Prediction Accuracy of End-of-Year Grade.

Units Completed	1st 20%	2nd 40%	3rd 60%	4th 80%	5th 100%
Correct Pos.	80%	45%	38%	33%	59%
Type I	7%	14%	14%	16%	18%
Type II	20%	55%	63%	67%	41%
Correct Neg.	93%	86%	86%	84%	82%

Table 4. Model Prediction Accuracy of Units Completed by Quintile.

However, a model that utilizes student work metrics across an entire year to predict end of the year grades is not as useful for informing interventions. Therefore, we repeated the analysis with models that only utilized a fraction of the data from the school year to determine their predictive validity of each of these models for end-of-year grade and units completed.

The full year of data was divided into academic quarters and then into sets of decreasing number of continuous quarters (1 through 4). Thus there are two 3-quarter subsets, Q1Q2Q3 and Q2Q3Q4, three 2-quarter subsets, Q1Q2, Q2Q3, and Q3Q4, and four 1-quarter subsets, Q1, Q2, Q3, and Q4. The model definition had to be adjusted accordingly to use more local measures of prior knowledge and work completion. Work completion was simply set to the total number of units completed during the sampled time. Prior knowledge was set to the grade earned in the quarter prior to the first quarter in the sample, or the grade from the prior year if the sample includes Q1.

The results of each regression are shown in Table 5. The diligence measure is significantly predictive of both end-of-

	End-of-Year Grade		Units Completed	
Samples	β (p-value)	R^2	β (p-value)	R^2
Q1	1.85(<.001)***	0.40	2.23(<.001) ***	0.43
Q2	1.69(<.001)***	0.45	1.72(<.001) ***	0.41
Q3	1.08(<.001) ***	0.45	1.16(<.001) ***	0.44
Q4	2.52(<.001) ***	0.51	2.09(<.001) ***	0.39
Q1Q2	1.93(<.001) ***	0.50	2.17(<.001) ***	0.52
Q2Q3	1.95(<.001) ***	0.56	2.09(<.001) ***	0.56
Q3Q4	1.11(<.001) ***	0.56	1.18(<.001) ***	0.55
Q1Q2Q3	2.11(<.001) ***	0.59	2.36(<.001) ***	0.63
Q2Q3Q4	2.04(<.001) ***	0.62	2.16(<.001)***	0.62

Table 5. Predictive validity over varying sample time windows.

Samples	1st 20%	2nd 40%	3rd 60%	4th 80%	5th 100%
Q1	59%	30%	10%	22%	41%
Q2	59%	35%	25%	17%	55%
Q3	55%	25%	25%	26%	59%
Q4	79%	48%	32%	39%	50%
Q1Q2	64%	30%	30%	22%	55%
Q2Q3	55%	20%	30%	26%	64%
Q3Q4	59%	10%	40%	35%	68%
Q1Q2Q3	64%	25%	25%	22%	64%
Q2Q3Q4	59%	30%	10%	22%	41%

Table 6. Model Positive Classification Accuracy of End-of-Year Grade.

Samples	1st 20%	2nd 40%	3rd 60%	4th 80%	5th 100%
Q1	65%	45%	29%	19%	27%
Q2	55%	35%	46%	19%	59%
Q3	70%	30%	42%	29%	55%
Q4	58%	40%	42%	38%	50%
Q1Q2	65%	45%	38%	19%	45%
Q2Q3	65%	45%	46%	14%	59%
Q3Q4	75%	40%	42%	38%	82%
Q1Q2Q3	75%	55%	42%	29%	59%
Q2Q3Q4	70%	40%	42%	29%	73%

Table 7. Positive Classification Accuracy of Units Completed

year grade and total curriculum units completed in a year across all time subsets.

The quintile analysis was repeated for each of the time subsets. The percent of correct positive labels was calculated for each dataset and averaged across all the datasets. The top and bottom quintiles of the End-of-year Grade regression models had a mean accuracy of 63.8% with mean standard deviation of 8.6%. The middle three quintiles of the End-of-year Grade regression models had a mean accuracy of 27.0% with a mean standard deviation of 8.7%. The top and bottom quintiles of the Units Completed regression models had a mean accuracy of 61.5% with a mean standard deviation of 11.3%. The middle three quintiles of the End-of-year Grade regression models had a mean accuracy of 36.1% with a mean standard deviation of 6.9%. Thus even without a full year of data, the model retains its prediction accuracy at all quintiles, though as can be seen in Tables 6 & 7, the model estimates begin to have larger errors as data size decreases.

Understanding the Diligence Measure

We followed this analysis with a partial correlation analysis to validate the relationship between our diligence measure and other SRL constructs. The partial correlation analysis included gender, ethnicity, and free and reduced lunch in the models. The predicted diligence measure using the full year of data is compared with the survey measures and the results are shown in Table 8.

Survey Measure	Correlation (p-value)
Math Interest	0.25(.01) **
Theory of Intelligence	0.05(.596)
Self-Efficacy	.258(.007) **
Mastery Approach	.284(.003) ***
Performance Approach	0.189(.051)
Performance Avoidance	.06(.52)
Effort Regulation	0.337(<.001) ***

Table 8. Partial Correlation with Diligence.

The diligence measure is significantly correlated with its analogous SRL construct, effort regulation, highlighting the predominant effect size of self-control on average during the usage of the tutor. There are also strong correlations with mastery goal orientation and self-efficacy ratings again supporting the hypothesis that these constructs moderate self-control. Likewise, both performance achievement goals were uncorrelated with diligence as anticipated. Furthermore, domain-interest is significant as expected because this is a domain-specific measure of self-control. Thus the agreement between the partial correlation analysis and theory bolsters the construct validity of this model.

Robustness to Data Sparsity

The robustness of the model to sparser data was tested through an initial analysis of a second data set from a set of 96 Geometry students from the same school and the same academic school year. The students in the geometry classes had about 5,000 transactions over the entire year, and thus had about 1/3 the data on average over any time-window compared to the Pre-Algebra dataset.

	End-of-Year Grade	
Samples	β (p-value)	R^2
Q1	.345(.17)	0.621
Q2	0.303(.43)	0.612
Q3	0.450(.31)	0.615
Q4	0.283(.18)	0.612
Q1Q2	0.259(.16)	0.618
Q2Q3	-0.618(.54)	0.612
Q3Q4	0.206(.46)	0.616
Q1Q2Q3	0.200(.26)	0.615
Q2Q3Q4	-0.391(.55)	0.612
Q1Q2Q3Q4	0.177(.30)	0.619

Table 9. Predictive validity over varying sample time windows using sparse samples.

The same ordinary least-squares regression was performed where the models included gender, ethnicity, free or reduced lunch, interest in math, and previous math achievement. All variables were normalized in order to facilitate interpretation of coefficients. The results of the regressions are shown in the Table 9. In this case the model shows that the smaller dataset was not significantly predictive of end-of-year grade at any time-window length.

DISCUSSION

In this paper, we introduced a model for measuring student diligence using online behavioral traces of an intelligent tutoring system. This method expands on an existing model by leveraging the characteristics of the intelligent tutoring system context to be able to draw inferences on quantity and quality of student effort. The result is a measure of diligent practice that has strong predictive ability on long-term academic outcomes even when only utilizing a relatively short time-sample. There is some initial evidence that the system needs a reasonably large sample of student activity in order to make more accurate predictions of long-term outcomes based on diligence measures. It remains uncertain based on the initial analysis conducted, whether this inaccuracy is because the data collected is sampled too infrequently to build an accurate picture of student diligence or because the student's measured diligence isn't reflective of an aggregation of all learning activities completed by the student.

The study found supporting evidence of how motivation and metacognitive measures such as achievement goals and self-efficacy influence diligence longitudinally. Interestingly, the higher predictive strength of this diligence measure at the extremes in contrast to the reduced predictive power at intermediate values is a result that is worth further investigation. Do intermediate diligence students have more varied academic exertion across academic activities? Are students in this range only measured as less diligent in the system, while they may tend to work more or less diligently on written homework or while studying for exams? Conversely, are extremely non-diligent and extremely diligent students more likely to apply constant effort and focus across all activities in the class? Is the varied diligence associated with academically relevant offline behaviors such as peer tutoring or frequently asking the teacher for help?

In this study, report card grades were utilized as prior knowledge measures, and therefore limited the extent to which smaller time-windows could be utilized to assess diligence. This leaves several open questions for future investigation. Many online courses do not span longer than a few months, and thus a diligence measure designed to identify potential low performers that requires 9 weeks of data is likely too slow to provide intervention support. Thus this model needs to be validated using alternative knowledge and progress measures that can be sampled more frequently. Knowledge tracing algorithms provide a much more fine grained picture of both prior knowledge and student learning, though they no longer capture learning gains from offline activities and thus may challenge the model accuracy. This is a promising avenue for future investigations into the robustness of the model to time-window length.

With an online behavioral measurement in hand, several new avenues of research can be opened. Especially in environments such as MOOCs where student motivation is a known problem [18], the proposed diligence instrument can be used to identify categories of low motivation and diligence students for more targeted study. Furthermore, interactions between diligence, self-regulated learning, and cognition can easily be explored through existing behavior-mining methods [19]. This instrument also creates opportunities to experiment at scale with a range of self-control interventions such as suggesting behavioral changes that alter the typical study context or scheduling [11], or encouraging more challenging learning activities [21].

CONCLUSION

This paper introduces a model that can measure student diligence unobtrusively through data generated when students interact with course materials. Furthermore, it can support more sophisticated research into the impact of various interventions on student diligence. Ultimately, the model can support the identification of patterns of diligent behaviors that lead to long-term academic success, uncovering a range of effective non-cognitive interventions and also elucidating the relationship between self-control based constructs, motivational states, and how micro-behaviors aggregate to produce specific long-term outcomes.

ACKNOWLEDGMENTS

The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education through Grant #R305B150008 to Carnegie Mellon University. We would like to thank Carnegie Learning, Inc., for providing the Cognitive Tutor data supporting this analysis. The opinions expressed are those of the authors and do not represent the views of the Institute of the U.S. Department of Education.

REFERENCES

1. Vincent Aleven, Ido Roll, Bruce M McLaren, and Kenneth R Koedinger. 2010. Automated, Unobtrusive, Action-by-Action Assessment of Self-Regulation During Learning With an Intelligent Tutoring System. *Educational Psychologist* 45, 4: 224–233. http://doi.org/10.1080/00461520.2010.517740
2. Carole Ames and Jennifer Archer. 1988. Achievement goals in the classroom: Students' learning strategies and motivation processes. *Journal of Educational Psychology* 80, 3: 260–267. http://doi.org/10.1037/0022-0663.80.3.260

3. Ryan Shaun Baker, Albert T Corbett, and Kenneth R Koedinger. 2004. Detecting Student Misuse of Intelligent Tutoring Systems. In *Intelligent Tutoring Systems*. Springer Berlin Heidelberg, Berlin, Heidelberg, 531–540. http://doi.org/10.1007/978-3-540-30139-4_50
4. Albert Bandura. 1994. Self-efficacy In VS Ramachaudran (Ed.) Encyclopedia of Human Behavior, 4, 71-81.
5. Roy F Baumeister, Brandon Schmeichel, and Kathleen Vohs. 2003. Self-regulation and the executive function of the self. In *Social Psychology Handbook of basic Principles* (2nd ed.). New York, 197–217.
6. Matthew L Bernacki and Steven Ritter. 2013. Hopewell 2011-2012. Dataset 613 in DataShop. Retrieved from https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=613.
7. Matthew L Bernacki, Timothy J Nokes-Malach, and Vincent Aleven. 2013. Fine-Grained Assessment of Motivation over Long Periods of Learning with an Intelligent Tutoring System: Methodology, Advantages, and Preliminary Results. In *International Handbook of Metacognition and Learning Technologies*. Springer New York, New York, NY, 629–644. http://doi.org/10.1007/978-1-4419-5546-3_41
8. Jeni L Burnette, Ernest H O'Boyle, Eric M VanEpps, Jeffrey M Pollack, and Eli J Finkel. 2013. Mind-sets matter: A meta-analytic review of implicit theories and self-regulation. *Psychological Bulletin* 139, 3: 655–701. http://doi.org/10.1037/a0029531
9. Angela L Duckworth and Laurence Steinberg. 2015. Unpacking Self-Control. *Child Development Perspectives* 9, 1: 32–37.
10. Angela L Duckworth and David Scott Yeager. 2015. Measurement matters assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher* 44, 4: 237–251.
11. Angela L Duckworth, Tamar Szabó Gendler, and James J Gross. 2016. Situational Strategies for Self-Control. *Perspectives on Psychological Science* 11, 1: 35–55. http://doi.org/10.1177/1745691615623247
12. Carol Dweck. 2000. *Self-theories: Their role in motivation, personality, and development*. *Psychology Press*.
13. Andrew J Elliot and Kou Murayama. 2008. On the measurement of achievement goals: Critique, illustration, and application. *Journal of Educational Psychology* 100, 3: 613–628. http://doi.org/10.1037/0022-0663.100.3.613
14. K Anders Ericsson, Ralf T Krampe, and Clemens Tesch-Römer. 1993. The role of deliberate practice in the acquisition of expert performance. *Psychological Review* 100, 3: 363–406. http://doi.org/10.1037/0033-295X.100.3.363
15. Stephen Fancsali, Matthew L Bernacki, Timothy J Nokes-Malach, Michael Yudelson, and Steven Ritter. 2014. Goal Orientation, Self-Efficacy, and“ Online Measures” in Intelligent Tutoring Systems. *CogSci*.
16. Brian M Galla, Benjamin D Plummer, Rachel E White, David Meketon, Sidney K D'Mello, and Angela L Duckworth. 2014. The Academic Diligence Task (ADT): assessing individual differences in effort on tedious but important schoolwork. 39, 4: 314–325. http://doi.org/10.1016/j.cedpsych.2014.08.001
17. Veronika Job, Carol S Dweck, and Gregory M Walton. 2010. Ego Depletion—Is It All in Your Head? *Psychological Science* 21, 11: 1686–1693. http://doi.org/10.1177/0956797610384745
18. Hanan Khalil and Martin Ebner. 2014. MOOCs completion rates and possible methods to improve retention-a literature review. *World Conference on Educational Multimedia*.
19. John S Kinnebrew, Kirk M Loretz, and Gautam Biswas. 2013. A Contextualized, Differential Sequence Mining Method to Derive Students' Learning Behavior Patterns. *JEDM - Journal of Educational Data Mining* 5, 1: 190–219.
20. Kenneth R Koedinger, Albert T Corbett, and Charles Perfetti. 2012. The Knowledge-Learning-Instruction Framework: Bridging the Science-Practice Chasm to Enhance Robust Student Learning. *Cognitive Science* 36, 5: 757–798. http://doi.org/10.1111/j.1551-6709.2012.01245.x
21. Kenneth R Koedinger, Jihee Kim, Julianna Zhuxin Jia, Elizabeth A McLaughlin, and Norman L Bier. 2015. *Learning is Not a Spectator Sport: Doing is Better than Watching for Learning from a MOOC*. ACM, New York, New York, USA. http://doi.org/10.1145/2724660.2724681
22. Carolyn MacCann, Angela Lee Duckworth, and Richard D Roberts. 2009. Empirical identification of the major facets of Conscientiousness. *Learning and Individual Differences* 19, 4: 451–458. http://doi.org/10.1016/j.lindif.2009.03.007
23. Gladwell Malcolm. 2008. *Outliers: The story of success*. New York: Little.
24. Walter Mischel, Yuichi Shoda, and Monica L. Rodriguez. 1989. Delay of gratification in children. *Science* 244, 4907: 933–938. http://doi.org/10.1126/science.2658056
25. National Research Council. 2011. *Assessing 21st Century Skills*. National Academies Press, Washington, D.C. http://doi.org/10.17226/13215

26. Paul R Pintrich. 1991. A Manual for the Use of the Motivated Strategies for Learning Questionnaire (MSLQ).
27. Paul R Pintrich. 2000. An Achievement Goal Theory Perspective on Issues in Motivation Terminology, Theory, and Research. 25, 1: 92–104. http://doi.org/10.1006/ceps.1999.1017
28. Steven Ritter, John R Anderson, Kenneth R Koedinger, and Albert Corbett. 2007. Cognitive Tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review* 14, 2: 249–255. http://doi.org/10.3758/BF03194060
29. Nicole Shechtman, Angela H DeBarger, Carolyn Dornsife, and Soren Rosier. 2013. *Promoting grit, tenacity, and perseverance: Critical factors for success in the 21st century*. Washington.
30. June P Tangney, Roy F Baumeister, and Angie Luzio Boone. 2004. High Self-Control Predicts Good Adjustment, Less Pathology, Better Grades, and Interpersonal Success. *Journal of personality* 72, 2: 271–324. http://doi.org/10.1111/j.0022-3506.2004.00263.x
31. Eli Tsukayama, Angela Lee Duckworth, and Betty Kim. 2013. Domain-specific impulsivity in school-age children. *Developmental Science* 16, 6: 879–893. http://doi.org/10.1111/desc.12067
32. Phillip H Winne, John C Nesbit, Vive Kumar, et al. 2006. Supporting self-regulated learning with gStudy software: The Learning Kit Project. *Technology …*
33. Barry J Zimmerman. 2008. Investigating Self-Regulation and Motivation: Historical Background, Methodological Developments, and Future Prospects. *American Educational Research Journal* 45, 1: 166–183. http://doi.org/10.3102/0002831207312909