Exploring Behavioral Measurement Models of Learner Motivation

Steven C Dang

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

February 26, 2022
CMU-HCII-21-109

Thesis Committee:
Ken Koedinger, Carnegie Mellon University (Chair)
John Stamper, Carnegie Mellon University
Geoff Kaufman, Carnegie Mellon University
Artur Dubrawski, Carnegie Mellon University
Sidney D'Mello, University of Colorado: Boulder

Submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy.

The research reported here was supported, in whole or in part, by the Institute of Education Sciences, U.S. Department of Education, through grant R305B150008 to Carnegie Mellon University. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education

Abstract

No learning happens until students make the choice to engage, regardless of how well-designed and personalized a lesson. While advanced algorithms have been developed to personalize and accelerate learning, similar highly quality models and algorithms aren't available for intelligent support of student motivation. The greatest challenge lies in the lack of high-quality measurement models to support the administration of motivational interventions. Existing models focus on the observable engagement behaviors of students. These measures are prone to noise from non-learner-specific influences as opposed to reflecting the underlying motivational drivers of engagement. This complicates the task of leveraging these analytics for assessing individual student motivational needs to support greater engagement.

In this dissertation, I address this gap through the development of a model to measure student diligence, their capacity to self-regulate and engage with learning activities. I leverage prior research in psychology and psychometrics to identify behavioral metric candidates. Through secondary analysis of a year-long longitudinal dataset of log data from students learning with intelligent tutoring systems, I evaluate the viability of these behavioral measures to estimate student diligence. I further develop these measures by leveraging theory to account for some cognitive, temporal, and social confounding factors. My analysis indicates that these behavioral measures, while better indicators of diligence, are still prone to other sources of noise that make the measures unreliable.

To address the unique challenges of measurement with observational data, I explore the viability of diversifying the model inputs by leveraging multiple operationalizations of diligence for estimation. I demonstrate that multi-operational models possess more desirable psychometric properties than any individual measure. Furthermore, I developed the Learner Engagement Simulator (LEnS) to generate data that reflects the challenges of estimating motivational constructs due to unobserved influences from the social and environmental context. My analysis of the simulated data reinforces the findings with real student data that multiple operationalizations of diligence increases the estimator accuracy.

for Judith,
whose love and support over the years has made this work possible.

for Amon and Matilda,
whose smiles and antics revive my weary mind.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Learning scientists and technologists continue to push the boundaries on building higher quality educational solutions, but ultimately, students must still make that choice to engage in order for learning to happen. Learning processes are an intricate dance between cognition and motivation [4], and effective instruction must adapt to both in order to more optimally enable learning. Advances in educational data mining and learning analytics have led to widely adopted models that can effectively trace student cognition and knowledge. However, engagement models focus on the moment-by-moment or overall superficial engagement of students instead of attempting to differentiate underlying motivations of students from contextual factors that may be driving engagement.

Assessing student qualities beyond domain skill and knowledge, such as student motivations, is becoming an increasing priority for practitioners, administrators, and policymakers [52]. However, tools for performing this measurement for use in instruction or administration are currently inadequate [63]. Past research in this field has used survey-based measures or specially designed behavioral tasks. The data collected from survey instruments are prone to many biases such as desirability and acquiescence bias, and behavioral tasks are prone to practice effects. Models derived from log-data generated through the usage of educational applications provide an opportunity to collect measurements of student motivations unobstrusively. Such measurement methods are not prone to the same biases as self-report and behavioral task measures, so for high stakes settings could complement such instruments to build a more valid compound measurement model.

Developing behavioral models of motivational constructs that are specific to the learning experiences of an educational technology product is a non-trivial task. Instead of attempting to identify and analyze behaviors that are specifically relevant for a particular construct, I approach this measurement challenge through the lens of engagement. This approach leverages existing work on learner engagement and engagement analytics to develop measurement models of motivation. I attempt to disambiguate latent motivations by leveraging details about the contexts where student engagement breaks down.

An overview is shown in Figure 1.1 of the many interacting factors that drive engage-

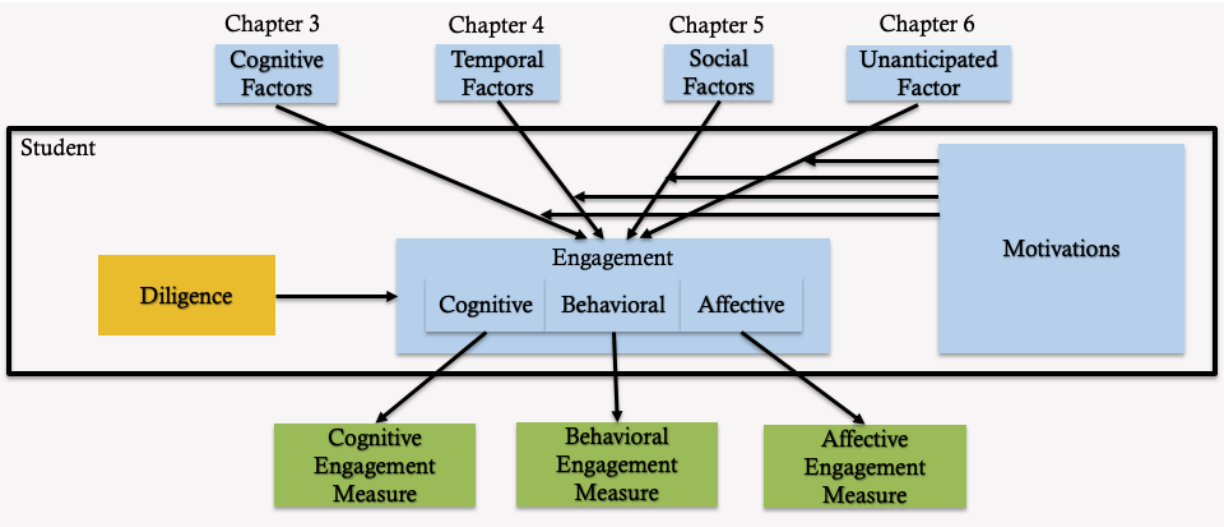

Figure 1.1: Model of measurement challenges with observational log data with overview of corresponding analysis chapters

ment and define the challenges of identifying diligence from measures of engagement. Prior educational data mining work has focused on the cognitive contexts of learning. Log data enables observation of the problem-solving process, making visible many cognitive factors that influence learners' success and behaviors. However, motivation is not strictly influenced by these cognitive factors. Student's aren't immediately able to focus at a moments notice, and their ability to maintain that focus tends to wane over time. Class time coming to an end can create pressures on students to perform. Despite the theoretical importance of these temporal factors in influencing behavior, models of learning do not encode these factors. Likewise, the social context can have a major influence on students' motivations. Educational psychology has focused on understanding the complexities of how individual student learning is driven by perceptions and expectations of peers, teachers, and family. Because these other people are not visible to the learning applications, this introduces a major hurdle in accounting for the role of these factors in driving student engagement. Furthermore, observational data, such as log data collected during learning with educational technology, is vulnerable to noise due to random unanticipated confounding factors, which may be cognitive, temporal, or social in nature.

In this thesis, I develop a measurement model of student diligence, their ability to self-regulate and focus on learning instead of more enjoyable alternatives. As shown in Figure 1.1, I explore in the next three chapters how to account for cognitive, temporal, and social factors evident when estimating diligence with log-data based behavioral measures of engagement. In chapter 3, I operationalize diligence as gaming the system [18], and I explore how interactions between gaming and challenge levels within the curriculum moderate the influence of motivation through a correlation analysis with survey measures. In chapter 4, I perform some feature engineering to encode in the data how long students have been working. I leverage this information to evaluate whether there is evidence that temporal effects predicted by self-regulation theory are evident in the log data traces. In

chapter 5, I develop some new features in the data that encode some facets of the social context. I explore whether the relative timing of students' engagement within a class session might carry information about motivations. In addition to developing diligence measurement models that factor cognitive, temporal, and social influences, I explore the use of multiple operationalizations within a measurement model as a approach to reduce the risk of bias from other unanticipated factors in the model. I evaluate the viability of this approach with both real student data and simulated data.

1.2 Motivated Decision-making

What goes on in the minds of students as they resist the urge to chat with friends to instead focus on completing the lesson activity? Why do some students manage to focus their attention onto the learning task assigned to them while other students give in to the desire to engage in other pursuits? This ability to self-regulate for academic pursuits, aside from intelligence, is one of the most reliable predictors of academic achievement [48]. What are the mechanisms that drive students decisions to stay on task?

Models of self-regulation have focused on the role of executive function to focus attention on the target task while also inhibiting desires to focus on other tasks. However, the relationship between executive function and observations of students' abilities in the classroom are very weak. Prior work has shown that the way learner's represent the task in their minds influences their effectiveness at inhibiting the undesired task [2]. Additionally, learners may employ strategies to manipulate their contexts, such as sitting far away from distractions, to reduce the need to inhibit the undesired task [72]. Therefore, the specifics of a particular context tend to have a greater effect on a learner's ability to self-regulate than the learner's innate capacity to inhibit and direct attention appropriately.

Motivation is defined as the orienting and invigorating impact on both behavior and cognition of prospective reward [62]. Students' motivation to self-regulate is linked to the nature of the tasks at hand. Value-based decision making research has demonstrated that this sort of self-regulation process can be represented as a decision between different possible actions where some subjective, weighted internal valuation of the outcome of each action informs the choice [76]. In these models, students choose the action that yields the the most personally valuable outcome. In academic contexts, value-based decision making models align with expectancy-value theory models that elaborate the specifics of decision making on learning tasks. In expectancy-value theory, learners apply their own subjective expectations of an outcome given a particular action and the associated values that might result from such an outcome [11]. The most common formulation for this relationship assumes a multiplicative relationship between expectancy and value. Given some universe of available actions A, for each action, a, there are k possible outcomes. For outcome, i, the student has some expectancy, $e_{a,i}$, of achieving the outcome, and a prospective value for successfully attaining the outcome, $v_{a,i}$. Student's then choose the action at time t, denoted a_t, according to equation 1.1 such that the action maximizes prospective reward over all possible actions and prospective reward for an action is the expectancy-weighted average of the value of all the possible outcomes for the action.

$$a_t = argmax_{a \epsilon A} \sum_{i=0}^{k} e_{a,i} v_{a,i} \tag{1.1}$$

A students' expectancy for a particular outcome given an action, their self-efficacy, are based on a subjective judgement informed by many factors including past experience with the task as well as domain and topic related confidence. Self-efficacy tends to vary across domains, topics, and skills [54]. With each experience, students update their expectations which will influence expectations in the future. Therefore the expectancy for the same action is expected to evolve over time as a result of these many influences[61].

Students attribute several types and quantities of task value to an outcome. There are four main task values associated with academic tasks [21]. Intrinsic value describes a value that students attribute completing the task itself due to the characteristics of the task such as an interest in the topic discussed in an assignment. Utility Value describes the utilitarian value gained from completing a task that services some goal such as performing well in the class or entering a particular career. Attainment value describes the importance of an outcome with respect to some aspect of an student's identity, ideals, or sense of competence [38]. Costs describe values that negatively impact a task valuation. Costs can be attributed to several factors including effort required for the task, effort put forth for other tasks, loss of alternative opportunities or value, and emotional costs [64]. For a particular outcome, student, the total value for the outcome is the generally treated as the sum of all of these components.

Task value is not considered to be constant. [7] found that children tended to have decreases in utility value for many academic subjects as well as intrinsic interest in reading over the course of several years in pre-adolescence. However, intrinsic value in math and music did not tend to change over that same period. [30] demonstrate the role that goals play in influencing task value dynamics over time. Achievement Goal theory models the effect of the types of goals that students hold [42, 10]. In this theory, goals have two independent attributes represented by a two-dimensional dichotomy. On one axis, goals can be either mastery or performance oriented. On the other axis, goals can be avoidance or approach. 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. The goals students hold influence the types of value they perceive for a task [60]. Performance goals are shown to be associated with utility value through a focus on task outcomes and subsequent academic performance [12]. Mastery goals are associated with both intrinsic value and utility value, through a focus on the task process as opposed to the outcome [34]. Students hold multiple goals simultaneously with varying degrees of strength [12]. Goals may vary both in type, number, and priority over time as a consequence of active self-regulation based on task feedback. Similar to expectancy, the task value associated with an outcome for a particular action are different as a result of the different contexts in which they are experienced [21].

In this model of motivated decision making, students have their own unique goals,

expectancies, and task values. These play a role in allowing the student to make a value-based decision about not only how to engage in a learning task, but also whether to continue to inhibit other activities to keep focused. During the learning process, students constantly self-regulate through a process of planning, action, evaluation, and reflection [41]. While expectancy-value can explain the planning and action phases, SRL research introduces different processes where students update expectancies, values, and goals based on an evaluation of the outcome of actions taken. An unsuccessful attempt will tend to drive down expectancies in the future[56]. Reflection on the completed task after updating expectancies may lead to realizations that a different action may be more appropriate in the future [8]. For instance, students may realize that help-seeking may be more valuable in situations of great difficulty. Similarly, students may realize that spending the time to attempt the difficult problem was not as rewarding as more leisurely and attainable activities such as watching entertaining videos. Some learners may reflect on this alternative outcome as being less rewarding for certain goals such as the desire to make an impression of diligence on the teacher. Making choices in the face of different short term versus long-term rewards is the role of self-regulation. In situations where learners identify this tension, they may re-prioritize their goals or define new goals so that choices may be more aligned with long-term over short-term rewards.

Academic Diligence is 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 [55]. The construct is an operationalization of self-regulation for the academic domain, and attempts to capture a trait-like capacity of individuals to self-regulate in academic settings. Measures of diligence have been shown to be align with conscientiousness, a personality trait linked to hard-work and perseverence, and is predictive of long-term academic outcomes. Though motivated decision-making is a process that is influenced by a set of motivation constructs that are potentially dynamic over time, evidence from personality-psychology indicates that behavior tends to be stable over time. Though the dynamics of how the hierarchy of constructs tends to shift over time to produce stable trends in behavior, current evidence supports the viability of this trait-like construct for predicting future behavior.

1.3 Measurement Models of Motivation

A significant body of prior work has focused on assessing moment-by-moment motivation through detectors of affect [33] and engagement [77, 66]. However, work analyzing the link between fine-grained behavioral measures and motivational goals and dispositions is much more limited. [40] created a rational model of student affect that leveraged a range of individual attributes including Big 5 personality measures and achievement goals. This work established the value of students' achievement goals on predicting moment by moment motivations as inferred by affect.

Several researchers attempted to identify task specific behaviors that rationally should be linked to achievement goals. [51] attempted to relate help-seeking behaviors while using an ITS to achievement goals. Researchers expected mastery-oriented students to be more

likely to use a glossary or index resource, while performance-oriented students might tend to ask for hints from the tutor instead. No significant relationship between self-reported achievement goals and help-seeking behaviors was found. However, task achievement goals as predicted by choice of help resources did relate to learning outcomes as would be predicted by achievement goal theory.

[59] expanded on this work and attempted to relate task choice, where descriptions of each task were closely linked to corresponding achievement goals, to self-reported achievement goals and learning outcomes. In this work, task achievement goals as inferred by task choice predicted learning outcomes for the lesson but did not align with self-reported achievement goals. However, self-reported achievement goals were more predictive of course outcomes. Researchers speculated that self-reported goals might reflect an average tendency to be motivated by particular goals over a range of tasks within the domain and thus explaining alignment with more aggregated measures such as course outcomes.

Gaming the system, a pattern of behavior where students abuse the design of the learning environment to answer a particular question, is a well-documented behavior that has been linked to poor learning outcomes [19]. In [29], the authors test the relationship between a range of student motivations and gaming the system behaviors across two different ITS's. The study results supported a link between gaming behaviors and some motivational measures but not others. One of the strongest results indicated that student's attitudes and interest towards the domain was related to observed gaming frequency. There was also strong support for a link between experiences of frustration and gaming as well as a lack of drive to motivate themselves on tasks in general as well as in the face of challenge. The results demonstrated mixed or weak support for a relationship with growth mindset and perceptions of the helpfulness of the ITS help resources. Interestingly, the researchers failed to identify a relationship between observed gaming and performance goals, though the performance goal measures were not drawn from validated achievement goal instruments. Furthermore, this study used strictly observed gaming frequencies. Subsequent work has identified the joint role of contextual and student factors in explaining gaming behaviors [39, 78].

Though not a model based on naturally observable behavior, [55] developed and validated a math-based digital behavioral task for measuring academic diligence. The task measured 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, two measures are collected, the total time spent solving math problems and the total problems solved. These measures were linked to conscientiousness, a personality traits associated with perseverance and working hard, and were predictive of many long-term academic outcomes. This simple task demonstrated the viability of developing a diligence measure based on the engagement choices that students make while learning, though challenges remain in overcoming the complexities of observational data.

1.4 Methodological Challenges

Past research in educational psychology has relied on survey-based measures to measure the expectancies, goals, and task values that learners have. These measures have been collected at different granularities including at the domain, topic, activity type, and task level. Unlike survey measures which are more portable across learning environments, operationalizing how these constructs manifest in fine-grained behavior for some specific learning environment is both necessary and non-trivial. There have been two main approaches to defining behaviors that might be indicative of motivational constructs, theory-driven and empirically-driven. Theory-driven approaches build on past research that describes the many factors that might interact with a target construct in influencing learner behavior. Model developers then have to consider the specifics of a target application and the target population to identify how factors from theory map to the learning context and predict behaviors within the universe of possible behaviors available to the learner. Empirically-driven approaches leverage ground truth labels collected using survey-based or observation-based methods in combination with digitally observable traces of the learner such as log data. Labels are then used with machine learning algorithms to define models of fine-grained behavior associated with the measured construct.

One common empirically-driven method for collecting labels for is the BROMP method [67]. In this method, coders observe every learner in a classroom for a brief time and record codes for the student's state. These codes are used as labels for supervised machine learning algorithms to develop fine-grained behavioral models from learner log data. This method has been used to build models for disengaged behavior [19], gaming the system behavior [19, 58], and affect [43]. There has been some work that has demonstrated that models trained over some content might generalize to other content within the same software [69]. Additionally, these models are as accessible to develop as the process for collecting the ground truth observation data. However, the data collection method is difficult to scale, and so may be biased by the particular sample of learners used to train the model. Research has shown that there are differences in the prevalence of different patterns of gaming the system behavior across the urban/suburban/rural divide [78]. Despite the great degree to which gaming-the-system models have been studied, development teams of new products cannot leverage these other models. Instead, teams must go through the effort of applying the BROMP method to collect data labels themselves, a significantly less accessible process for incorporating such models into a product.

Researchers have also used experience sampling to collect self-report labels for developing models. [73] demonstrated the viability of this method for developing automated detectors of mind-wandering by probing users pseudo-randomly during a learning activity to reflect on whether they had mind-wandered. This method is significantly more accessible than the BROMP method as it can be performed through the application the user is already using, and so it will readily scale. As a self-report measure, the data labels are prone to many of the issues common with self-report measurement [9]. In assessment settings where students know their responses might be used to support teacher monitoring of behavior, student responses are prone to social-desirability bias [63], providing responses that match the socially desirable state of diligent work as opposed to reporting mind-

wandering. For measurements of constructs such as intrinsic value, student responses may reflect reference bias [13], where students may have different standards for what degree of interest in math is associated with a three out of five.

Another method uses human coders to develop computational models through a more labor intensive qualitative analysis process. [23] demonstrate this method on cognitive tutor data, and [49] demonstrate this method in an open-ended learning environment for learning scientific inquiry. In this class of methods, a subset of log data is selected for qualitative analysis. Trained human coders apply qualitative coding methods to review short sequences of logs for individuals and apply labels to the data according to their coding dictionary. This class of methods shares many characteristics with the BROMP method except this method is less prone to sampling bias because label data can be created for any population that the application currently serves. However, the method relies on a human coder's ability to operationalize and interpret constructs in the coding dictionary in terms of the limited log data available. The significantly smaller degree of information available to human coders to make judgements raises some questions about accuracy. [28] demonstrated that for gaming the system behaviors, expert coders were able to label data with good reliability, but the degree to which these results apply to other motivational constructs or over data collected from different applications is still an open question.

Another bottom-up method requires collecting labels at significantly lower resolution than in BROMP or experience sampling methods. [50] demonstrated a method for identifying patterns of behavior associated with effective self-regulated learning processes while using an open-ended science learning environment. In this work, they leverage achievement data at the level of learners as a supervision signal to be used in conjunction with a sequence mining algorithm to identify patterns of behavior that differentiate low achievement learners from high achievement learners. Given the very rich space of possible behaviors and strategies available to students in the open-ended learning environment, these methods demonstrated a capacity to automatically identify learner behaviors from fine-grained log data associated with a desired training signal collected at the learner level. For trait-like constructs such as good/poor self-regulation ability, where the construct can be assumed to be constant over the data collected, this method can work quite well. However, these measures are much more sensitive to overfitting and are prone to biases from unobserved confounds such as cultural differences in how students tend to work [78, 68] or classroom specific technology usage patterns [65, 46].

As opposed to these bottom-up approaches, which allow models to be defined from the data, top-down methods primarily depend on theory to define models. [22] demonstrate an a-priori thresholding process for operationalizing SRL theory into a measurement model on fine-grained data. The authors applied SRL theory in a cognitive task analysis to develop a model consisting of a set of if-then-else rules representing a decision tree model for solving problems with help-seeking. By leveraging theory from prior research, the model benefits from the greater likelihood of a more generalizable model. In order to apply this model to the data, the model operationalized concepts such as "Familiar at all?" and "Sense of what to do?" using a set of calculated values in the data and thresholds that are set to values that the authors describe as "intuitively plausible, given our past experience". This a-priori heuristic is difficult to reproduce and requires an intuitive sense of how users

interact with the system. Alternatively, the values might be set using the data itself, but then the model may face similar threats of overfitting if generalized to new environments or content.

Another example of the top-down approach is demonstrated by [32], where they extended a model of off-task behavior [18] to a narrative-centered learning environment. This open-ended environment involved actions such as navigating a character around a virtual world, interacting with objects, and talking with non-player characters. In lieu of a usable measurement model, the authors developed an alternate operationalization of off-task behavior that required insight into the pedagogical value of possible interactions in the game. [32] were motivated to develop an off-task behavior measurement model informed by the findings of [18], indicating that student learning is negatively impacted by such behaviors. Despite appearing to match the [18] construct on its face, validation of the model indicated that the model developed by [32] was not measuring the same construct. There was no significant relationship between learning and observed frequency of the behavior, and no correlation was found with survey based motivational measures either. The authors mention in the discussion that the operational definition captured a broad behavior that in different observable contexts, an observer might draw not label the behavior as being off-task. This example of operationalizing a model demonstrates some of the main challenges in translating existing theory and models to new contexts and applications. There is ultimately some degree of interpretation that must be performed by the model developer in translating a model to a new learning environment. This requires both defining the behaviors and the contexts that affect the interpretation of the behavior. These problems are evident in the broader problems of a lack of generalizability of models in the space of motivation-related learning analytics research [84] and highlights an opportunity for greater support.

Top-down approaches trade the benefits of more generalizable models for more labor costs from learning science experts. Alternatively, bottom-up approaches are able to more automatically discover behavioral models, but such models are subject to validity threats due to the methodology that training labels are collected, and over-fitting to latent characteristics of the sample population. There is an opportunity to explore the middle ground between these two types of approaches with multi-operational models. Multi-operational models can leverage the benefits of top-down models by leveraging theory to inform multiple operationalizations of a construct, while pairing with bottom-up approaches to train system-specific fine-grained behavior models.

Chapter 2

The Dataset

The work in this dissertation primarily uses a dataset drawn from Datashop [35] that was collected as part of a year-long study [47] in a suburban middle school in a mid-atlantic state. The students used the Carnegie Learning Cognitive Tutor software (CogTutor). A screenshot of a student using the algebra 1 version of the software is shown in figure 2.1. The CogTutor software provides adaptive instruction based on a fine-grained skill representation of the domain. The application divides problems into steps that must be answered individually and each map to independent skills in the domain model. Student practice problems are selected according to whether they have demonstrated mastery of necessary skills. The instruction is also scaffolded, allowing students to request multiple levels of hints at every step of the problem, providing on-demand problem scaffolding that provides increasingly informative support to the students. The data logs generated by the software are transformed into the standard learning data format specified by [35] before being utilized in this analysis. This format specifies how long students spend on every interaction, whether the action was correct, incorrect, or a hint, and what skill is associated with a specific problem step. Each interaction is represented as a single student transaction in the dataset, which includes over 4M such transactions across all students observed.

The dataset includes recorded learning transactions of students using the tutor approximately two class-periods per week for a full school year. The dataset also includes demographic, achievement, and motivational survey measures. The demographic information collected includes gender, ethnicity, free-or-reduced lunch status, and special education status. Achievement data includes end-of-year grades from the prior academic year as well as the grade for each academic quarter and a cumulative end-or-year grade for the year transaction data was collected. Survey measures were collected at the beginning and the end of the course to measure students' motivational goals and dispositions. Each scale utilized was drawn from well-validated instruments. Survey measures include scales for interest in math [37], self-efficacy [24], effort regulation [3], growth mindset [86] and achievement goals [10]. Responses for each scale were summed to represent students' motivation along each dimension. The specific questions and their respective response scales are referenced in appendix C.

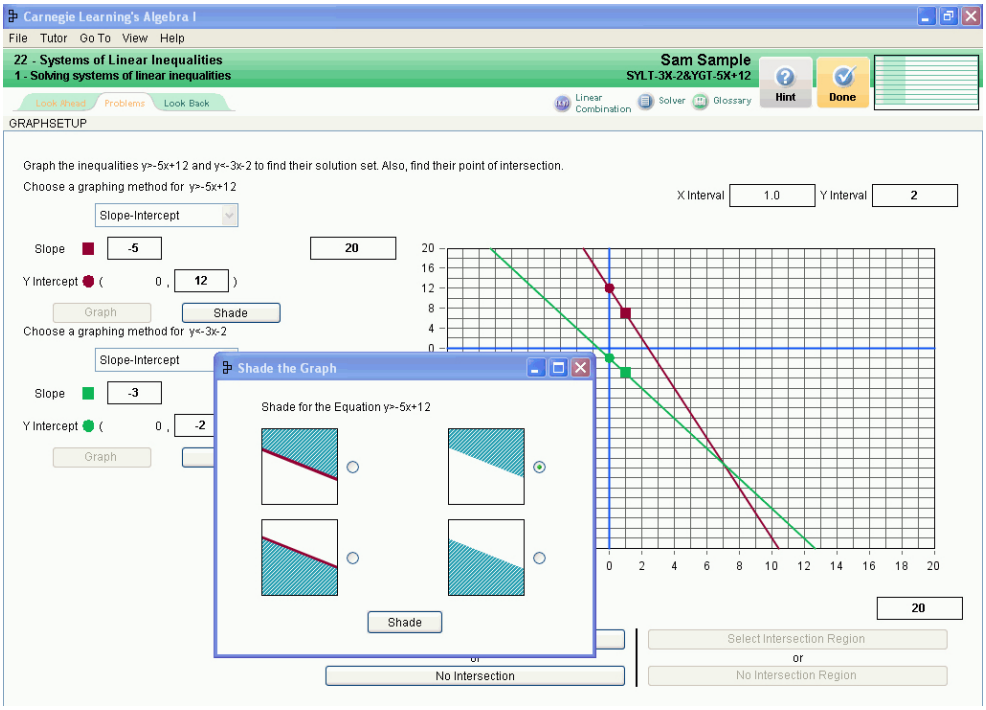

Figure 2.1: Screenshot of Algebra 1 CogTutor demonstrating the multi-step student input interface design that enables fine-grained problem-solving process tracing

The dataset includes a total of 426 students, but not all students have complete or valid data that is usable for analysis. An overview of the different subsets of the sample that had complete or partial data are shown in top row of table 2.1. Student courses are identified by relating the unit names of the problems solved to a dictionary of expected unit names associated with each course curriculum. There are 426 students in the total dataset where each student in that dataset with data from at least one category of data, transaction, demographic, achievement, or survey data. There were also 155 students that had no accompanying demographic or external achievement information ("Only TX" column). 107 of these 155 students belong to classes where there is at least one student with demographic or achievement information, so the transactions for these students are included for some of the subsequent analyses in this document. Details are elaborated in the corresponding chapters. Transactions for the other 48 students ("Unknown Classes" row), are excluded from all analyses. Of the remaining 271 students, 45 of these students ("Incomplete" column) were found to have missing or invalid data for some achievement or motivational survey measures. The subsequent analyses conducted in this document are performed using this set of 226 students.

One important characteristic with this dataset is that while there are 426 students in the dataset, many of these students are missing some form of transaction, demographic, achievement, or survey responses. The missing data is not random across the dataset. The students are spread across three courses, Pre-algebra, Algebra 1, and Geometry. Most of the missing data is from students in Algebra 1 courses, with 127 out of 147 students having

Table 2.1: Data exclusions impacting sample size by course

	All	Only TX	Incomplete	Complete
Total	426	155	45	226
Pre-Algebra Classes	127	2	17	108
Algebra 1 Classes	147	104	23	20
Geometry Classes	104	1	5	98
Unknown Classes	48	48	0	0

missing achievement data. While Pre-algebra and Geometry classes are between 9 to 25 students in size with a mean close to 16, these classes only have between 1-3 students being excluded due to missing data. Algebra 1 classes are similarly between 7 and 23 students in size with an average of 16. However, the vast majority of students are lacking achievement information where each class has between 1 to 4 students with complete data for analysis. It is unclear why data collection was poor for Algebra 1 classes, however, the students with complete data are included for most analyses unless otherwise stated in the study methodology.

From the full set of demographic information collected, a set of 271 students, the population includes 125 Pre-algebra students, 103 Geometry students, and 43 Algebra 1 students. The students are 48% male and 52% female. The population is from a suburban school in the mid-western United States where the students are predominantly white with only 2% of the population identifying as non-white. 23% of students receive free-or-reduced lunch, and 14% of students are identified with special education needs. These exact demographics vary slightly from the specific sample with complete data used in subsequent analyses, but exact breakdowns of the distributions are included in appendix A.

Chapter 3

Cognitive Factors

3.1 Overview

The contexts when students demonstrate failures to self-regulate their learning behaviors can be informative of their motivational goals [15], their perceived value of the activity [21], and their beliefs about their self-efficacy [17]. When aggregating behaviors such as time-on-task as in the academic diligence task, it becomes difficult to account for the contextual factors that may be influencing measures of student diligence. Unlike in the behavioral task where there are tight controls on the context that students complete the task, in observational settings these contextual factors they may be quite variable over the time the data is collected. Operationalizing diligence in a more fine-grained manner can enable better insight into the influence of contextual factors on student decisions. The study in this chapter explores the feasibility of leveraging observations of students' self-regulation as measured by gaming the system behaviors to measure student diligence. Furthermore, it seeks to answer two main research questions.

Research Question #1: How does the relationship between gaming and measures of motivation differ when gaming estimates are derived from either raw observations of gaming or using random effects models that account for both student-level and contextual variation.

Research Question #2: How does student performance on educational content with varying degrees of gaming frequency relate to their different motivational goals and dispositions?

3.1.1 Gaming the System Behavior

Gaming behaviors are defined using the heuristic model introduced by [58] as this model appeared to produce better kappa on unseen data from across multiple systems including the CogTutor. Using this model, individual transactions were labeled according to a taxonomy that captures a range of relevant behaviors such as thinking before a hint request, spending time reading hint requests, and variations of guessing behaviors. Transactions are labeled as gaming if they are a member of a set of subsequent transactions that matches

one of the thirteen expert identified heuristic patterns [58]. The patterns encode two primary types of gaming: guessing and hint abuse. Guessing patterns include placing the same answer incorrectly into multiple available answer slots and answering the same question rapidly with very small changes in the answer across attempts. Hint abuse patterns include not stopping to think about multiple subsequent errors before requesting help and rapidly requesting hints to seek a bottom-out hint, which in the CogTutor environment is simply the answer to the problem step given as the second or third hint. Transactions are rolled-up into student steps, where each student step encapsulates metadata about all the transactions associated with a problem step until a correct answer is reached. Each student step is labeled as gamed if any transaction associated with the step was also labelled gamed. The resulting student step data was utilized to calculate student and content gaming frequencies. Details of the implementation of the detector model are included in appendix B.

3.2 The Dataset

The analysis in this chapter uses the dataset described in chapter 2. This analysis utilizes a subset of the full dataset, including the 206 students from the pre-algebra and geometry courses with complete and valid observed tutor transactions, demographic information, achievement data, and motivational survey responses. Algebra 1 students were excluded from this analysis because there were inadequate students to gauge differences in content-level difficulty. The dataset included 3.5% gamed student steps. These numbers align reasonably well with gaming frequencies observed in prior work on CogTutor data. [19] found students gaming the system about 3% of the time based on in-classroom human observations. [58] found a slightly higher overall gaming frequency of 6.8% in their dataset utilizing the same detection model as used here. However, this deviation is not so different that it is due to significant unobserved differences in the populations.

The CogTutor content is organized hierarchically into multiple units. Each unit consists of several sections that themselves have multiple skills to be learned. Each section has problems that are divided into highly granular steps which each are associated with at least one skill. I grouped observations at the section level to capture differences across the curriculum with sufficient resolution while having sufficient observations across students to make reasonable estimates of gaming frequency. The data included 237 sections with a mean gaming frequency of 1.95% and a standard deviation of 1.7%. A number of sections were found to have no observed gamed steps, while the highest observed frequency was one section with 32% gamed steps.

Additionally, some curricular sections were excluded due to having low observations in the data. Transactions from 41 sections are excluded from the dataset because they were observed with less than 6 students completing any work in the section. These sections are excluded because such sections might be measurements of only the fastest working or highest achieving students, thus introducing a bias to observations of gaming within those sections. Furthermore, the data is divided into two subsets, hard sections and non-hard sections. Hard sections are defined based on how often students are observed gaming on

problems in that section. An 80% quantile threshold of 3.23% observed gaming frequency for each section separates the hard and non-hard sections. There were 156 non-hard sections and 40 hard sections. Additional descriptive statistics of observed gaming in each subset of the data can be reviewed in appendix A.5

Unlike in prior work, [19], no students were found to have never gamed throughout the year. The average student was observed gaming 3.66% of the time with a standard deviation of 1.16%. The minimum observed gaming frequency for students was 1.98% while the maximum observed was 11.95%.

3.3 Methods

In this study, four estimates of student gaming are generated and compare to each motivational measure using partial correlations controlling for gender, ethnicity, and free/reduced lunch status.

$$\theta_{ObservedGaming} = \frac{x_{NumGamed}}{N_{TotalSteps}} \tag{3.1}$$

$$P(Gamed) \sim (1|Student) + (1|Section) \tag{3.2}$$

$$\theta_{Gaming} = e^{\theta_{student}} \tag{3.3}$$

To investigate RQ1, student gaming tendency is calculated using only raw observations for each student as shown in Eq 3.1. Gaming tendency is estimated by fitting a model to predict gaming on each problem-step using a random effects model with a random effect for student and tutor-section as shown in Eq 3.2. The model is fit over all observed student steps and the student gaming tendency is found by calculating the exponential of the fitted random intercept, $\theta_{student}$, for each student as shown in Eq 3.3. To investigate RQ2, I analyzed estimated student gaming on the hard and non-hard section data subsets. Again, the random effects model in Eq 3.2 was used for each data subset to estimate student gaming.

For RQ1, I expect student gaming estimates from the random effects model to better correlate with motivation relative to observed gaming because the model takes into account variance in gaming due to sections, which may not be observed for all students, as well as accounting for statistical noise due to sampling of a rare event.

For RQ2, I investigate the hypothesis informed by design principles of psychometric behavioral tasks. Measuring a targeted construct requires straining the resource and identifying a metric upon which to differentiate subject performance. Therefore, I expect estimates of student gaming using only highly gamed sections will have a more significant relationship with motivational variables compared to data without highly gamed sections.

3.4 Results

The results of the partial correlation analysis are shown in Tables 3.1 and 3.2. The first row of both tables present evidence contrary to the results from [3]. Prior research found correlations with math interest, effort regulation, and growth mindset using only averages of observed gaming. However, in this dataset, only interest in the subject is related to gaming behaviors, and no other motivational measure has a significant correlation with student's gaming frequency.

On the other hand, the second row reflects correlations with student gaming estimated using a random effects model fitted with all of the data. In general, more motivational measures are correlated with these gaming estimates than those derived from the raw observations, which supports the hypothesis for RQ1. Comparing these results to [3], there are no direct measures of frustration, however it is possible that self-efficacy mediates whether student's experience of frustration explaining the correlation. Growth mindset is found to be marginally significant, which further bolsters the previous mixed evidence for a link between mindsets and average student gaming.

There are two cells where these correlations do not seem to agree with prior research. Effort regulation is expected to be correlated both as a matter of face validity as well as because prior research found a relationship between gaming and students' drive to persevere on academic work.

Table 3.1: Correlations between Gaming and Motivation Measures

Data Subset	Math Interest	Self Efficacy	Effort Reg.	Growth Mindset
Observed	-0.22**	-0.10	-0.11	0.00
All	-0.17*	-0.16*	-0.11	-0.14(.)
High Gaming	-0.19*	-0.14(.)	-0.10	-0.11
Low Gaming	-0.16*	-0.19*	-0.16*	-0.14*

(.) - $p < 0.10$, * - $p < 0.05$, ** - $p<0.01$, *** - $p<0.001$

Table 3.2: Correlations between Gaming and Achievement Goals

Data Subset	Mastery Approach	Performance Approach	Performance Avoidance
Observed	-0.03	-0.01	-0.05
All	-0.20**	-0.10	-0.15*
High Gaming	-0.14(.)	-0.08	-0.11
Low Gaming	-0.25***	-0.14(.)	-0.21**

(.) - $p < 0.10$, * - $p < 0.05$, ** - $p<0.01$, *** - $p<0.001$

The link between achievement goals and gaming are mixed. In [29], the authors assessed performance goals using questions such as, "If you had your choice, what kind of extra-credit projects would you most likely do". It is unclear how this question maps

to achievement goals, however, performance approach goals are not significant as might be extrapolated from prior work. On the contrary, mastery approach and performance avoidance goals are correlated with gaming. This relationship is rationally derived from the theory on self-regulation and motivation, but not predicted by specific prior work. Overall, the random effects model yielded a significant relationship to more motivational constructs than gaming estimates from raw observations.

The results from estimating gaming using only highly gamed sections, the third row of each table, are contrary to what is expected. Many of the correlations that appear when using all of the data, are weakened or not significant when using only the hardest questions. While the loss of significance with some constructs could be an artifact of random sampling from the full dataset, this does not explain the results seen in the bottom row. When estimating gaming using only non-highly gamed sections, correlations arise with every available motivational construct as seen in the fourth row. This is an unlikely consequence of sampling from the population and supports the idea that student gaming performance on highly gamed questions is introducing additional noise to the available signal in the rest of the data. Thus, the evidence points towards student gaming behaviors in the non-highly gamed sections as being more informative of student motivations than behaviors in the highly-gamed sections where self-regulation is under greater strain.

3.5 Discussion

In this study, I demonstrate that leveraging random effects models to cope with statistical noise in observations of student's tendency to game on any given section better estimates student's gaming as related to their motivational goals and dispositions. Additionally, I provide initial evidence towards a measurement model of student's motivational goals and dispositions by leveraging observations of gaming. Results indicate that student gaming tendencies can be used to estimate diligence by accounting for the cognitive factors that may vary with the type and difficulty of problems that are encountered during practice.

Several correlations with gaming estimates appear contrary to prior research and merit further analysis. The significant correlation with both mastery approach and performance avoidance disagrees with the results found by [29]. This disagreement could be due to the independence of achievement goals from each other, where gaming may be driven by an aggregate motivation of all achievement goals. More analysis is necessary to bridge this seeming contradiction and understand how patterns of gaming across problems of varying difficulty and prior experience might support an interpretation of gaming as indicative of different achievement goal profiles.

Gaming frequency was leveraged as a proxy measure for a range of unencoded difficulty factors. While this includes factors such as poor classroom instruction or a poorly designed cognitive model, it also encapsulates difficulty of individual problem-steps. A natural next step would be to investigate how more detailed student skill models might improve estimates of perceived difficulty and corresponding enrich model understanding of the nuances of why students are gaming and how this relates to different motivational goals.

Prior work has also shown that on longer time-scales, motivational goals are not neces-

sarily stable [53]. In this study, I looked for relationships between pre-course motivations and in-course gaming behaviors. For students with fluctuations in achievement goals or self-efficacy, the contexts in which such students tend to game or disengage from the lesson in other manners might similarly change. Further analysis is necessary to investigate whether variations in gaming over time are similarly reflective of variations in motivational goals and dispositions over time.

Furthermore, the study included a fairly large body of students, but the observations were still limited to a single school in a particular region of the country with limited ethnic and socio-economic status diversity represented in the sample. Such factors are known to be correlated with variations in the types and frequencies of gaming behaviors observed in the population [78]. As such, I exercise caution in extrapolating these relationships beyond this demographic group without further validation.

Nonetheless, the results presented in this work lay the groundwork for further investigation into measurement models of motivational goals and dispositions that leverage an understanding of the contexts that strain students' self-regulation. Such unobtrusive measurement models hold the keys to a future where schools can better utilize instructional time that is currently occupied by standardized test and test-specific preparation while still receiving the student, and class-level performance measures necessary to support continuous improvement.

Chapter 4

Temporal Factors

4.1 Overview

Many teachers can relate to the struggle of keeping an entire class engaged as the end of the day approaches. Some students may be listening raptly while other have started packing their belongings. Many teachers use class management techniques, such as specific activities in the beginning of class, in anticipation of the difficulties in ramping up the engagement of the entire class [16]. Student motivation appears to vary systematically over the course of a class period. Many good teachers adapt to this reality. It seems appropriate that intelligent tutoring systems should as well.

Student procrastination, the failure to engage in a task in a timely fashion, has a well-established link to student motivations [27]. The nature of the tasks that students have difficulty engaging can be revealing about their individual goals [15], their perceptions of the value of the task [21], and their beliefs about their abilities to complete the task [17]. Similarly, the context of what drives students to quit can be equally telling about the same facets of student motivation [20].

Measures of quitting and procrastination leverage the easily observable dichotomy of student engagement, but are there other within-task student behaviors that might similarly indicate motivation? Quitting and procrastination are evidence of students' failure to exercise their self-regulation. In these moments, students are failing to direct their attention towards a less desirable but beneficial learning task, and instead opting to engage in more desirable non-learning tasks. Applying this self-regulation lens, it may be possible to understand student motivation by identifying and analyzing other observable moments during student work where students engage in less desirable behaviors for learning.

For instance, solving an extra credit problem on the homework may likely push the student's grade from a "B" to an "A" for the year. However, the problem will likely take an hour to solve and the student may have to skip soccer practice to find time to complete the problem. Observing the student's choices and behaviors in these critical moments of self-regulation can reveal student's underlying motivation. Prior models of self-regulated learning behavior have focused on the cognitive facets of a given task: its difficulty level [80, 81], its domain topic [57], its time cost [64], and its expected value to the student

[11]. However, research on self-regulation point to temporal factors that influence decision making.

Task switching research indicates that the exercise of self-regulation imposes a cognitive cost. Once an individual chooses to engage in a task, they do not always appear to be applying themselves with full effort [36]. Additionally, when a person is forced to change tasks rapidly, they are not able to perform at the same level as those given more consolidated spans of time to perform on the same task [14]. These studies imply that students are likely to perform at a reduced capacity when initially beginning work to perform on a task upon initially beginning work,

Ego-depletion models of self-regulation posit that the ability to regulate attention over time may tend to deplete as some time-driven function of an internal and limited resource [26]. Thus, motivation may also tend to wane over time leading to an eventual failure to self-regulate.

This chapter explores whether these temporal properties of self-regulation are evident in student behaviors through patterns in observations of their failures to self-regulate.

4.2 Related Works

Measuring self-regulation related constructs is not a new concept in the intelligent tutoring system literature. Prior work has developed a range of models for detecting self-regulation related behaviors.

4.2.1 Off-task Detection

Some of the earliest work in this space identified off-task student behaviors by identifying large gaps of time between interactions in the log data of student interactions [19]. Inferences on student skill improvement, in addition to whether the students asked for help or attempted a problem correctly/incorrectly following a long gap between interactions determined whether students were off-task while idle.

[75] developed models of mind-wandering, when students' attention and thoughts move off-task, which enabled detection of off-task behavior over much shorter time spans. These models leveraged information from videos and human labels of short time segments to train a supervised model to classify when mind wandering occurs. The features fed into the model included a range of low-level image processing features, facial features, inferred emotions, and temporal features that describe the dynamics of facial features and emotions during a short time interval. [79] extended this work given user self-reports of mind-wandering and included body position information.

4.2.2 Persistence and Quitting

[80] developed a model of student persistence by analyzing patterns of behavior that included observed student actions contingent on properties of the problems being worked and the student's skill on those problems. In this work, two types of students emerged,

where the authors posited that trait level differences in students' capacity for sustained attention lead to differences in learning strategies and persistence during problem solving.

[70] designed a game-based measure of trait level persistence and validated the measure against other existing survey and standard psychometric behavioral tasks. The measure looked at average time on unsolved versus solved problems given a wide range of difficulty levels.

In [81], the authors built models of quitting an educational game. They leverage many features including features of each level of the game, the current state of game progress of the student, and the time in the current level. The final model that emerged from the supervised machine learning process were focused around actions of the student and the state of progress and counts of actions at each level across and within attempts at the level, thus not including any of the limited temporal features given at model training time.

[57] attempted to predict when students would quit reading a given passage. In this work, the authors used semantic features of the reading passages, the recent context of what passage is being read, which passages have been read recently, and both current page and total reading time. Total reading time, a similar proxy to ego-depletion, was found to be a significant contributor to models of quitting with respect to the first page of a passage. The authors also implicitly investigated the role of task switching by predicting quitting at the beginning of a new passage compared to some other new page within a passage. While some of the data supports a differential impact of task switching and quitting, the authors do not explicitly explore how quitting behaviors vary over time.

4.2.3 Gaming the System

With intelligent tutoring systems that provide scaffolding supports through progressively informative hints and feedback, another behavior tends to arise called "gaming the system" [18]. These behaviors have been identified using information about a series of recent actions such as time spent or the number of recent hint requests and errors, and the characteristics of the problems worked, such as problem section and difficulty in those interactions [78]. Extensive work has attempted to determine what drives gaming behaviors. While some initial work determined that problem context better explained gaming behaviors over trait-like individual propensities to game [25], later work presented the opposite result using a different intelligent tutoring system [39]. A large multi-environment analysis was conducted that compared the types of gaming behaviors observed across urban, suburban, and rural contexts using three different intelligent tutoring systems [69]. The study found that across tutoring environments, students displayed different predominant gaming behaviors, which implies that the lure of certain types of gaming may be different given tutoring environment or problem-type affordances. Similarly, within tutoring environments, students from areas of different population density (eg: rural versus urban) display different predominant patterns of gaming. These differences point to how variation in work environment may have differential anticipated costs to gaming, while the variation within environment but across geographic regions point to possible cultural and thus motivational differences.

4.2.4 Research Question

Prior work has developed extensive models of self-regulation behaviors that demonstrate the importance of cognitive, contextual factors, and local temporal factors for influencing student's self-regulation decisions. However, these models have not investigated how self-regulation behaviors might vary systematically over time and how such trends relate to student learning. In this chapter, I am looking to investigate whether the within-session temporal properties of self-regulation are evident in student behaviors and whether these temporal trends are predictive of similar negative impacts on student learning.

Models of the cognitive cost of task switching imply that self-regulation related behaviors such as gaming the system are more likely to occur in the beginning of a work session. Similarly models ego-depletion imply that self-regulation related behaviors such as gaming are more likely to occur after students have been working for some time. I propose to investigate whether models of task-switching and ego-depletion are evident in some changes over time of the probability of gaming the system, a behavioral instance of self-regulation. I then investigate whether lower cognitive engagement as predicted by task-switching theory co-occurs with gaming the system. I follow this with an analysis to determine if failures in self-regulation during critical time periods are indicative of session-level motivation.

4.3 Methods

4.3.1 Dataset

The analysis in this chapter uses the full set of 226 students from the dataset described in chapter 2. In order to see temporal patterns, data was excluded from short sessions with length in the bottom 5% of all student session lengths, which was determined to be about 5 minutes. This excludes data from 494 sessions. The resulting observed student sessions ranged from 5 minutes to 58 minutes, with a median length of 32 minutes.

To measuring gaming the system behaviors, the same gaming model applied in chapter 3 is applied here. Details of how the gaming model is implemented can be review in appendix B. Overall the dataset consists of 3.5% of steps as being labeled as gaming behavior, where the majority of students are labeled as gaming between 3.1 to 4.5% of all observed steps with a minimum of 2.0% and a maximum of 13.4%.

4.3.2 Aligning Session time

In this chapter, sessions are defined as the working session of a student starting when they are first observed working to the time they stop working. Sessions are encoded in the original transaction data with a unique session identifier(session ID), so temporal patterns are inferred by batching student transactions by session ID and re-orienting time-stamps relative to the start of the session. The start of the session is defined as the time of the earliest transaction sharing a particular session ID minus the duration of the first transaction. The end of the session is the time of the last transaction sharing the same session ID.

One difficulty in measuring ego-depletion with observational data is in controlling for differences in the depleting effects of context. In ego-depletion studies, the task is controlled for and thus can be ruled out to explain observed differences in behavior. In intelligent tutoring contexts. The adaptive instruction will provide variably challenging and types of content and may differentially deplete students across the experiences within the same period of time. To overcome this issue, we leverage the insight that when two students begin working, they might be in similar states relative to their internal thresholds for self-regulation. We also assume that when two students stop working, they are in comparable states. If these two students stop working at different times, it implies similar start and finish attention states, but different depleting effects of context that were experienced over time. In order to account for these differences in uncontrolled contextual factors, we created an additional time measure that aligned individual student transactions within sessions by the percentage of the session time that has elapsed. This alignment facilitates comparison of transactions relative to the start and end of a session, scaled to the session length.

4.3.3 Modeling The Effect of Time

Theories of self-regulation imply different models of the effect of time on self-regulation. Attentional shift models posit a cognitive cost of task switching. These costs may cause some tasks to seem more difficult near the beginning of a session. Ego-depletion models imply a reduction of a limited capacity to self-regulation resource over time. These models suggest students may eventually find it difficult to continue in a task and signs of fatigue, such as gaming, may be revealed by an increased tendency to engage in gaming behaviors before finishing working. To test these model implications, we compare five random effect logistic regression models to determine how self-regulation may vary over the course of a session.

We introduce M4.1 as the baseline model for comparison whether any temporal models are more significantly more predictive than current best practices as suggested by prior gaming research. This model, includes random effects for both student and curricular section to control for the previously established impacts of student and context on student's tendency to game. The remaining four subsequent models similarly control for student and contextual factors while introducing alternative factors representing temporal effects.

To define the remaining four models, time is represented along two dimensions. In the first dimension, time is represented as either time elapsed since the student began working or percentage of total working time elapsed, as described section 3.2. Time elapsed models represent the default model informed by both ego-depletion and task switching theories. Percentage of time elapsed models test the hypothesis that such a representation better captures motivation as temporally relative to the most informative moments of student behavior. In the second dimension, time is represented linearly or quadratically. Linear models allow only one main temporal effect to be captured by the model, either a constant increase or decrease in motivation over the course of a session. Quadratic models can capture different effects at the start and end of the session that differ from each other and the middle of the session. All temporal variables are normalized over the full dataset for model interpretation.

M4.1: Baseline – Baseline model for comparison controlling for differences in student's tendency to game and contextual factors across curricular sections, such as average difficulty, that influence gaming.

$$Gaming \sim (1|Student) + (1|Section) \tag{4.1}$$

M4.2: Linear Session Time – Extending the baseline model M4.1 by adding a linear term for time-elapsed since the student has begun working

$$Gaming \sim TimeElapsed + M4.1 \tag{4.2}$$

M4.3: Linear Percent Time – Extending the baseline model M4.1 by adding a linear term for proportion of session time elapsed as a percentage of total time observed working.

$$Gaming \sim PctTimeElapsed + M4.1 \tag{4.3}$$

M4.4: Quadratic Session Time – This model extends model M4.2 by adding a quadratic term

$$Gaming \sim TimeElapsed^2 + M4.2 \tag{4.4}$$

M4.5: Quadratic Percent Session Time – In addition to the random effects in Eq 4.1, this model tests the hypothesis that students self-regulation resources are

$$Gaming \sim PctTimeElapsed^2 + M4.3 \tag{4.5}$$

4.4 Results

Table 4.1: Comparing models student gaming behaviors over the course of a work session

Model	BIC	AIC	LogLik
M4.1	434741	434703	-217348
M4.2	434682	434632	-217312
M4.3	434668	434619	-217305
M4.4	434454	434392	-217191
M4.5	454503	434441	-217215

The results of fitting each of the five models are shown in Table 4.1, including model performance as assessed by AIC, BIC, and log-likelihood. In general, all models with temporal factors outperform the baseline model, M4.1. This implies that temporal information has a significant effect on student's self-regulation behaviors. Additionally, both quadratic models, M4.4 and M4.5, are significantly better than their linear counterparts ($\chi^2 = 179$ ($p<0.001$) for M4.2 vs M4.4, and $\chi^2 = 242$ ($p<0.001$) for M4.3 vs M4.5). Likewise M4.4 and M4.5 are significantly better than baseline with $\chi^2 = 315$ ($p<0.001$) and $\chi^2 = 266$ ($p<0.001$) respectively. This supports the interpretation that there are non-monotonic differences in gaming the system behaviors between the start, middle, and end.

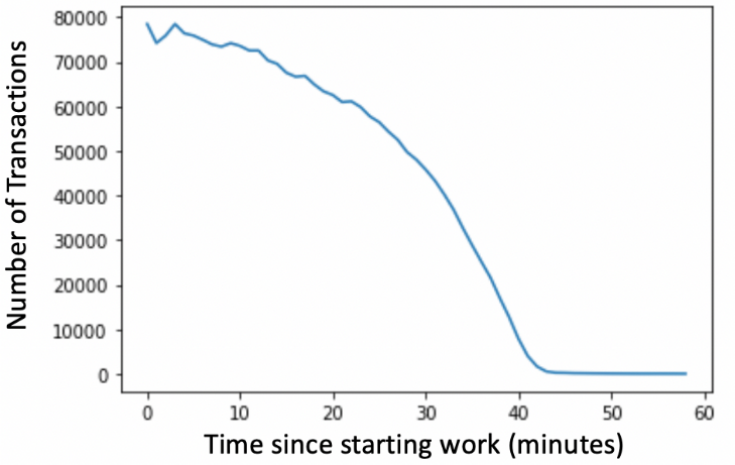

Figure 4.1: Volume of data observed over time in a working session

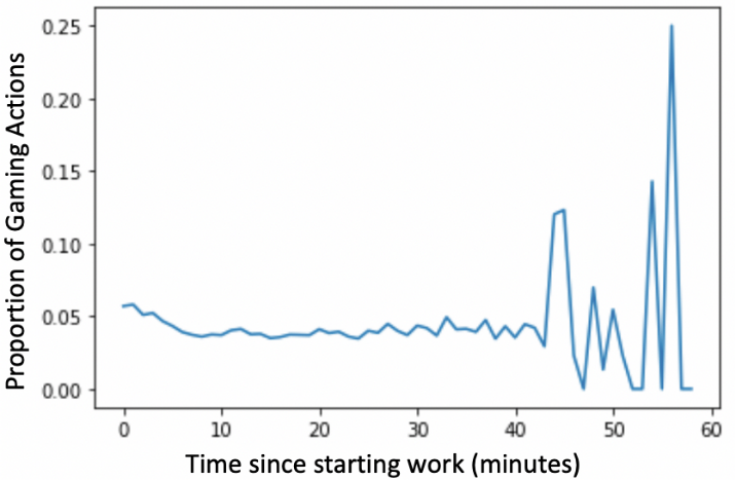

Figure 4.2: Proportion of gaming actions observed over time in a student's working session

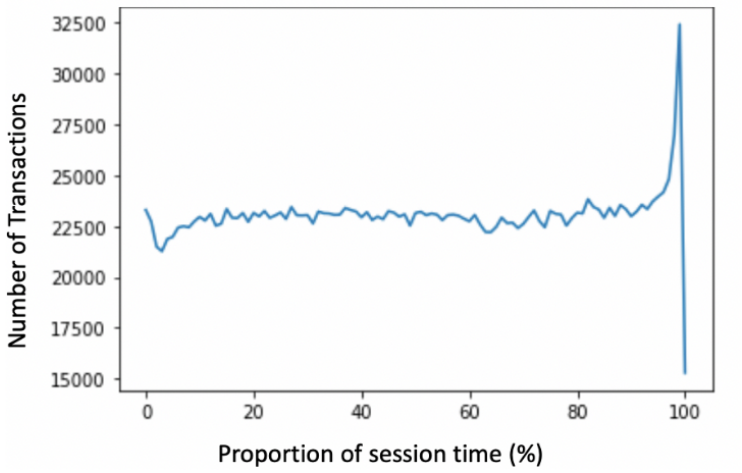

Figure 4.3: Volume of data observed over a working session aligned by proportion of overall session length

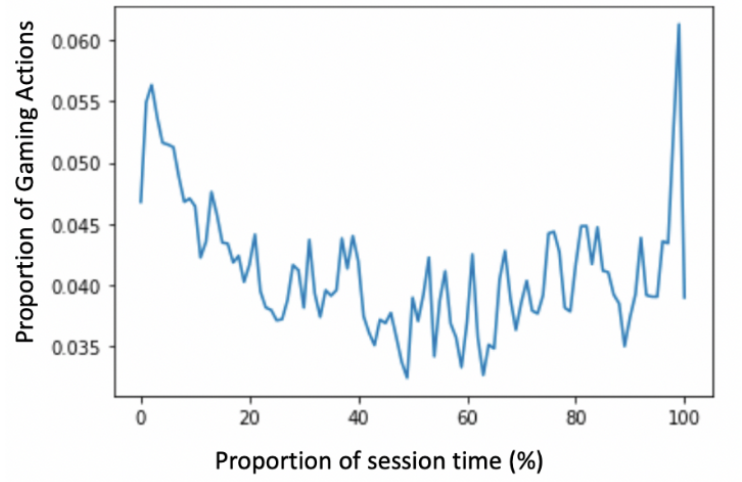

Figure 4.4: Proportion of gaming actions observed over time in a student's working session aligned by proportion of overall session length

Exploratory plots of proportion of gaming the system transactions over the session support these interpretations. Figure 4.2 and 4.4 plot the proportion of transactions identified as gaming the system behaviors across the session over minutes passed or proportion of total session time respectively. As expected from the quadratic fit models, each figure shows an increased proportion of gaming behaviors near the start and end of sessions.

A closer look at the data in Figure 4.1 reveals that there is a large student participation drop-off near the 43 minute mark. While whole class sessions seem to regularly measure about 60 minutes, students' login and logout times are quite staggered such that 99% of observed student sessions are less than 43 minutes in length. Only 82 out of more than 9800 sessions are observed where students worked continuously for between 43 and 60 minutes. Furthermore, analyzing gaming averaged over each minute of the hour, Figure 4.2, shows that this dramatic reduction in data is associated with very large and volatile estimates of average students gaming per unit time. Because of the low amount of data observed in the last 17 minutes of sessions longer than 43 minutes, it is hard to draw stronger conclusions about whether students are much more likely to display gaming behaviors if they are able to stay on task longer than 43 minutes, or if the volatility is due to random sampling bias.

A closer inspection of data in Figure 4.3 also shows some peculiar variability in data at the start and end of sessions. Because session time is divided evenly across the proportion of sessions, there is no a-priori reason to believe students have more or less frequent transactions at any time in the session. The small decrease in quantity of transactions near the start of sessions implies students take longer on average to complete actions near the start of work. The large spike of activity near the end implies students are taking less time per action shortly before stopping work. In both cases, the data sparsity issue seen in Figure 4.1 is not likely driving the changes in proportion of gaming seen in Figure 4.4. The small decrease in activity near the start is associated with the start of a broader downward trend in proportion of gaming behaviors that continues even after activity frequency flattens. The sudden increased frequency of transactions near the end of sessions is associated with a comparable spike in prevalence of gaming the system behaviors. However, because some gaming behaviors are defined by rapid actions in succession, this relationship is expected.

Taking the model comparisons and exploratory data analysis together, this evidence supports the interpretation that there are non-monotonic differences in gaming the system behaviors between the start, middle, and end of sessions.

Term	M4.4 - ß	Term	M4.5 - ß
Intercept	-4.215	Intercept	-4.217
Percent time elapsed	-0.265	Time elapsed	-0.283
$Percenttimeelapsed^2$	0.231	$Timeelapsed^2$	0.252

Table 4.2: Model coefficients for M4.4 and M4.5

Comparing the two quadratic models, M4.4 is the best fit model by all 3 measures, BIC, AIC, and Log Likelihood. The model details can be seen in Table 2. The variance in gaming attributable to curricular sections is 0.87. This translates to average gaming attributable to tutor context level factors to range between 0.23% and 8.4% for 95% of sections. The variance attributable to students is much smaller, 0.088. This translates

to average gaming attributable to trait-level student factors to range between 0.82% to 2.57%. An inspection of the model coefficients shows that the model predicts the average gaming level at the start of a session, $P(gaming|t = 0)$, is 4.1%. Average gaming at the end of the session, $P(gaming|t = 60minute)$, is 18.7%. The quadratic model reaches a minimum observed gaming of 1.3% at 23 minutes into the session.

An 18.7% average probability of gaming after working for 60 minutes appears to be very high given that gaming only occurs overall in the dataset in about 4.5% of all actions. As discussed in the previous exploratory data analysis, the very high gaming proportion observed in the last 17 minutes of sessions is potentially related to the increased volatility created from estimates drawn from small amounts of data. These estimates spike upwards as high as 25%, which corresponds with the dramatic difference between start and end gaming predicted by M4.4. Therefore, the model is reflecting this same artifact of the data.

Inspecting M4.5, the model predicts that gaming is more likely in the start and end of the session. The average probability of gaming decreases to 1.35% by the time the student has worked 67% of the total time. According to the model, we are 3.34 times more likely to observe students game the system near the start of work than near their peak level of focus. Likewise, it is 1.32 times more likely to observe gaming the system in the moments shortly before students stop work. This model appears to make less dramatic predictions that are more inline with expectations based on overall average frequencies of gaming while not reflecting the same uncertainties as M4.4.

These results support the hypothesis that self-regulation processes have an impact on the average occurrence of gaming the system behaviors over the course of a work session. Students in this data appear to experience decreased motivation near the start of work as would be predicted by the cognitive costs of task switching. Likewise, students appear to show some decreased motivation before stopping work as predicted by ego-depletion theories.

Gaming Indicates Cognitive Effort

If students are not observed to game the system early in a session, we expect that student motivation is likely higher around this time despite the brief slightly negative impact of task switching. This greater motivation allows students to bring greater cognitive resources to the work relative to days when gaming is observed near the start. When comparing assistance rates in the beginning of a session, the proportion of questions either answered incorrectly or with a request for help on first attempt, a student who is more cognitively engaged should be less likely to make errors or ask for help. Likewise, similar patterns should be associated with assistance rates near the end of students work.

We compared the assistance rates for sessions where a student is observed gaming in the first 10% of the session time (the first 3 minutes for the median session) to assistance rates where no gaming is observed in the first 10% of the session time. To calculate the assistance rate, the raw student transactions are aggregated by problem-step. The outcome of each step is determined by the first attempt at the step. The step is labeled as gaming the system if any of the aggregated transactions are labeled as gaming. Because patterns

of gaming generally involve either incorrect or help-seeking behaviors, steps that were labeled as gaming the system are removed before calculating the proportion of incorrect and help-request steps to overall steps observed in the portion of the session.

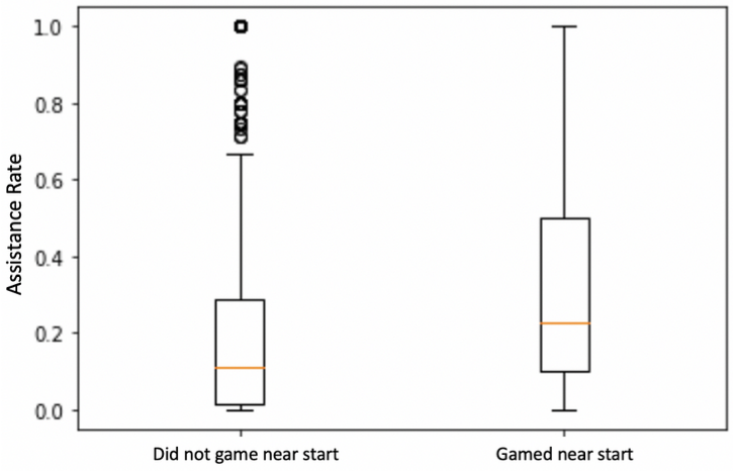

Figure 4.5: Comparing assistance rate at the start of sessions

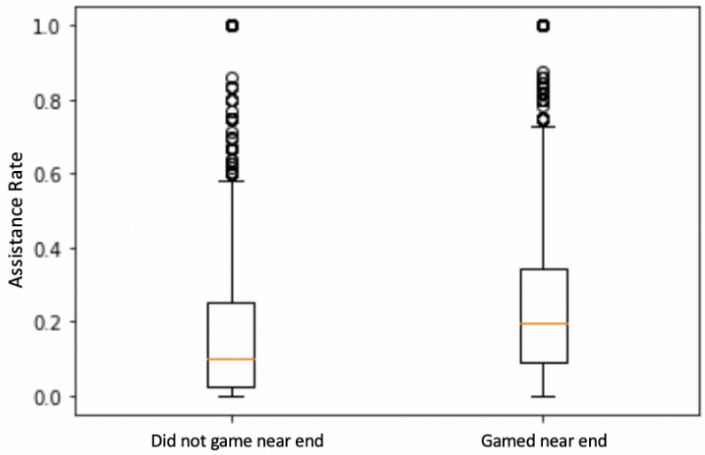

Figure 4.6: Comparing assistance rate at the end of sessions

The assistance rates in the start of sessions are shown in Figure 4.5 and were found to be significantly lower ($t = -15.22$, $p < 0.001$). The average assistance rate where gaming is observed is 30% ($sd = 25$) while the average rate when gaming is not observed is 21% ($sd = 26$). Similarly, Figure 4.6 shows boxplots for assistance rates in the last 10% of sessions. Rates were found to be to be significantly lower ($t = -11.6$, $p < 0.001$) with the average session where gaming is observed having a rate of 25.3% ($sd = 22$) compared to the average non-gaming session having a rate of 18.6% ($sd = 24$).

This simple analysis does not take into account factors such as question difficulty. It is possible that if students are working on difficult content near the start, then they are more likely to make errors and request hints. It also implies that more challenging material may impact how students evaluate the likelihood of prospective reward given their perceived abilities. This may lead students to believe that applying effort is unlikely to result in experiencing the reward or attempting to apply effort may have greater depleting effects that impact future actions. In either case, it is possible that more challenging material instead of task-switching or ego-depletion explains the relationship between increased assistance score and gaming behaviors near the start and end of work. However, these tests do provide compelling evidence for a possible impact of decreased cognitive engagement on some practice opportunities that can inform future modeling work.

Gaming Indicates Motivation Levels

Student's day-to-day average motivation level is affected by factors in the school, in the classroom, and in the student's life more broadly. A death in the family, fight with a significant other, or poor grade in another class might be weighing on a student's mind while that begin working. These broader factors may have a more general negative effect on student's ability to self-regulate during work, effectively directing all attention to the learning task at hand. If this is the case, these factors will act in combination with the additional impacts of task-switching or ego-depletion at the start and end of the session to impact a student's capacity to self-regulate. Thus, observing gaming the system behaviors at the start or end of a session may also be informative about a student's more general motivational level. In this section, we analyze gaming behaviors throughout the session using information about whether students gamed at the beginning or end of a session to improve predictions of gaming in the rest of the session.

Gaming at the start and end are defined the same as in the previous section. In the data, 29.7% of sessions are observed with gaming at the start while 32.0% of sessions have gaming at the end. Together 49.9% of sessions have instances of gaming the system in the start or end, while only 11.8% of sessions are observed with gaming in the start and end of the session. While gaming near the start or end might be indicative of session level motivational impacts, in this analysis we test whether seeing any gaming at the start or end is sufficiently informative or if start and end are differently informative.

To perform this analysis, we use the best previous model , M4.5 the quadratic percent-time-elapsed model. This model will control for the variance due to student and tutor contextual factors, removing concerns about confounds such as gaming at the start may be due to generally more difficult material that makes gaming more likely throughout the session. We compare models that add main effects for whether gaming was observed at the start or at the end as well as linear and quadratic interaction effects. The models are elaborated as follows:

M4.6: Baseline Quadratic Model – the baseline model from Section 4 analysis for comparison.

$$Gaming \sim PctElapsed + PctElapsed^2 + (1|Student) + (1|Section) \tag{4.6}$$

M4.7: Gaming at start/end main effect – M4.6 with a binary indicator variable of whether gaming is observed near the beginning of the session and a binary indicator variable of whether gaming is observed near the end of the session

$$Gaming \sim M4.6 + gamed_{start} + gamed_{end} \tag{4.7}$$

M4.8: Combined Gaming at start or end main effect – M4.6 with a binary indicator of whether gaming is observed at either the beginning or the end of the session

$$Gaming \sim M4.6 + gamed_{start+end} \tag{4.8}$$

M4.9: Gaming at start and end with linear interactions – M4.9 elaborates on top of M4.7 adding linear interactions with time.

$$Gaming \sim M4.7 + gamed_{start} * PctElapsed + gamed_{end} * PctElapsed \tag{4.9}$$

M4.10: Gaming at start and end with quadratic interactions – M4.10 elaborates on top of M4.9 adding interactions with quadratic time terms.

$$Gaming \sim M4.9 + gamed_{start} * PctElapsed^2 + gamed_{end} * PctElapsed^2 \tag{4.10}$$

Comparing M4.7 and M4.8, we see that including separate main effects for gaming at the start and gaming at the end leads to better models rather than combining the information into a single indicator of whether there were any self-regulation failures at either the start of the end of the session. This particular result is worth further investigation to understand how and why self-regulation at the start of a session is differently indicative of student motivation levels compared to gaming at the end of the session.

Table 4.3: Comparing models work session gaming given observed gaming near the start/end of a session

Model	AIC	BIC	LogLik
M4.6	434441	434503	-217295
M4.7	422316	422403	-211151
M4.8	427322	427397	-213655
M4.9	419913	420045	-209958
M4.10	418266	418402	-209122

The results in Table 4.3 indicate the best fit model is M4.7, the model with start/end gaming information and interactions with linear and quadratic terms. This model is significantly different from the baseline quadratic model ($\chi^2 = 49.42$, $p < 0.001$) and establishes the informativeness of gaming in the start or end of a session on student's motivation levels through the time that students are working. Details about the model is given in table 4.4.

The variance accounted for by section and student level random effects are reduced in comparison to the baseline quadratic model reported in Section 4. The variance attributable to student factors was found to be 0.0789, which translates to an average gaming level of 0.64% to 1.91% for 95% of students. The variance attributable to section level

Table 4.4: Coefficients for start/end gaming with quadratic interaction terms

Term	β
Intercept	-4.489
PercentTimeElapsed	1.129
$(PercentTimeElapsed)^2$	-1.251
Gamed at start	0.301
(Gamed at start) * (PercentTimeElapsed)	-1.480
(Gamed at start) * $(PercentTimeElapsed)^2$	1.170
Gamed at end	0.356
(Gamed at end) * (PercentTimeElapsed)	-0.490
(Gamed at end) * $(PercentTimeElapsed)^2$	0.900

factors was found to be 0.7527, which translates to an average gaming frequency of 0.20% to. 5.79% for 95% of sections. This implies that a significant fraction of observations of gaming that were previously explained by section-level factors appears to now be explained by motivational factors indicated by gaming at the start or end of a session.

Table 4.5: P(Gaming) Main effect predictions given start/end gaming observations

Context	Game (t=0)	Game (t=opt)	Game (t=100)	Start Odds	End Odds
No Gaming start or end	0.35%	1.43%	0.21%	0.24	0.15
Start Gaming	2.14%	2.14%	0.66%	1	0.31
End Gaming	0.18%	2.10%	1.71%	0.086	0.81
Start + End Gaming	53.1%	1.72%	5.1%	30.9	2.98

Table 4.5 contains the predicted gaming attributable to the main effect terms in model M4.10. The first column describes average predicted gaming at the start of work. The third column describe average predicted gaming at the end of work. Because the model includes quadratic terms, the second column is included to describe the optimum (minimum or maximum) probability of gaming throughout the session. The fourth column describes the odds ratio the chance of gaming at the start relative to the optimum point. The fifth column describes the odds ratio of the chance of gaming at the end compared to gaming at the optimum point. The complexity of the model can make the model challenging the interpret, however there are some important trends indicated by the model. If gaming is observed only in the start of a session, gaming is most likely to occur similarly near the start and will reduce over the course of the session as evidenced by the odds of gaming being greatest at the start relative to the end. Likewise, observing gaming only at the end of the session implies that students tend to be well regulated near the beginning of the session and will appear to fatigue over the session until near the end where the odds fall slightly. When students are not observed gaming at the start or end, there is a corresponding low probability of observing gaming near the start and end. However, over the course of the

session, the model predicts that these students become more likely to have slightly reduced motivation until the latter half of the session where attention on the time pressure of the end of class might increase motivation through the end of class. In the limited sessions where students are observed gaming at the start and end, the model predicts a much greater propensity to game throughout, with a 53% chance in the start and a 5% chance near the end.

Taken together, these results support the conclusion that gaming at the start and end of work are indicative of session-level motivational factors influencing student behavior. It also provides initial evidence for separable constructs indicated by gaming at the start versus at the end. Each of these constructs appears to have different degrees of impact on underlying student motivation factors and the resulting decision processes that lead to observable behaviors.

4.5 Discussion

We have treated gaming the system behaviors as indicators of student's self-regulation. Task switching and ego-depletion theories of self-regulation predict a temporal pattern to student's abilities to self-regulate over the course of a class period. Predictive model comparisons are supportive of the hypothesis that both task switching and ego-depletion are evident in the patterns of student behaviors over each class session. Further analysis indicates that observations of self-regulation behaviors in the start and end of class might be indicative of both temporally immediate degrees of cognitive engagement as well as more session or day-level influences on motivation.

Open questions remain about how student models could operationalize task switching or ego-depletion. The work presented, uses information about the full student session to represent time, though such information is not available to real-time models. This raises the question of how should student's prior behaviors inform a predictive models of student ability to task switch or ego deplete? To what degree do students display consistency in their ability to task switch quickly or manage ego-depletion more effectively across sessions? Over the course of months or years? To what degree are these capacities independent or can correlations be attributable to other latent motivational causes?

We believe these findings highlight the importance of leveraging student models that incorporate temporal variables in the design of learning activities. Problem selection algorithms may want to be biased for lower challenge or greater interest to overcome negative effects of task switching. Similarly, activities may want to incorporate changes in the rhythm of the activity in order to periodically re-engage student attention as it wains over time. This work exposes an unexplored design space for how educational activities could incorporate temporal effects of student motivation to better enable student learning.

In this work, we introduce the importance of considering temporal factors in addition to content-related cognitive factors to more effectively support students' motivational trajectories within a work session. These findings extend the rich body of work on modeling student motivational and cognitive processes with self-regulated learning. Students are not machines, and they do not always jump immediately into tasks full throttle or have

the endurance to work as long as they are asked. Hopefully, a future that recognizes these dynamics can take intelligent tutoring systems one step closer to emulating the capabilities of effective teachers.

Chapter 5

Classroom and Social Factors

5.1 Overview

Understanding student motivation through engagement requires understanding the choices that students are making. Online learning environments offer the ability to directly observe student's thoughts about the problem-solving process at very fine-grained levels. However, students' choices to engage in a learning activity are not strictly informed by the characteristics of the task itself. Self-regulation is by definition a choice. While students are working on problems, many of the relevant choices can be inferred through knowledge about the current problem state. However, task-specific choices are not the only important factors that are driving student engagement. As in the academic diligence task, students are balancing engaging with tutor with alternative activities unrelated to the tutor. Students are working in a classroom with a teacher imposing demands on their behavior to work on the tutor during some time and nearby friends that may be doing something else that is entertaining. In order to better estimate student's capacity to self-regulate through time-on-task measurement, it should be important to contextualize these observations with respect to the classroom context to better understand the balance of the choices students are making.

One common format for classes using an intelligent tutoring system is that a class will meet in a dedicated computer lab where each student is able to log into their own individual computer. At the start of class, students are expected to login and continue working through the entire class session. At the start of class, students must transition themselves away from whatever they are doing to begin working on the tutor. Students may be talking with friends, working on some non-class related task, or just rolling thoughts over in their mind. Different students may tend to have different rituals as they transition into the class, depending on the friends that may cross their path, or whether their class schedule typically has assigned homework that might be due later in the day. The ability of students to transition into working on the tutor will be dependent on their current task, their perception of value of that activity, and their motivational values and expectations for succeeding on the work in the tutor. Each scenario has some value for the student relative to working on the tutor. Depending on the types of motivations that drive the student,

they will have some associated ability to transition away that is measurable by how long they take to transition. Typical socio-cognitive motivational constructs such as self-efficacy, intrinsic interest, and achievement goals play a role in this decision to engage, but they only impact the prospective reward of one choice. Students are ultimately balancing choices of when/how to engage with the tutor against some alternate activity and the relative values of these two activities impacts the probability that the student chooses to transition at any given time.

Student engagement time is influenced by two major factors in the classroom. The first is the amount of time given to the students to work by teachers, which provides an implicit normative expectation for how much working time is expected of students. Teachers dedicate a certain amount of class time as opportunity for students to work on the tutors. In turn, this time creates an implicit standard for students to meet, which creates an extrinsic incentive for students to engage with the tutor that is dynamic depending on how the student has managed their time. Students are therefore likely making engagement decisions some of the value of engagement is a function of what proportion of the provided opportunity the student had worked. The second factor influencing students is their social context. While teacher's provide some implicit expectation about time to work, the activity of peers also influences a student's decision to engage. Even if the students are all in a classroom where the teacher expects them to be working, if most peers are not working, this decreases the incentive for a student to work because the normative expectation is that the student is performing on par with others. The inverse should also be true, if most students are working, this should increase the perceived incentive for engaging with the tutor over alternative activities. Therefore, the relative engagement of students to in-classroom peers should also impact students' decisions.

The goal in this chapter is to investigate whether there is value in leveraging social and classroom contextual information for improving estimation of student diligence. In particular, I am seeking to investigate two research questions.

Research Question #1: Is diligence estimation improved by accounting for how students are utilizing the opportunity provided by teachers?

Research Question #2: Does taking into account students utilization of opportunity relative to their peers improve diligence estimation?

5.2 Methods

5.2.1 Dataset

The analysis in this chapter uses the full set of 226 students from the dataset described in chapter 2. Unlike in chapter 4, sessions in this chapter are defined as inferred class sessions as opposed to student working sessions. Building on the student sessions investigated in the previous chapter, class sessions are inferred using data from students identified as belonging to the same class through a shared class ID in the data. The specifics of inferring

class session are described in the next section.

5.2.2 Measures of Diligence

One of the most common measures of student diligence is time-on-task. This direct measure of student engagement is an easily accessible measure that can be inferred from log data. Naturally, not all logged time is engaged time, and there has been prior work [19] that developed models to differentiate disengaged from engaged logged-in time.

One challenge with leveraging this raw measure of time-on-task is that it is unclear how much time each student was able to work. A student working 45 minutes in a classroom where the teacher provided three 60-minute class periods to work is significantly less engaged than a student working 45 minutes given only a single 60-minute class period in that same week. The opportunity that teacher's provide students to work is a valuable reference for contextualizing the observed time that student's are working.

Given only logged interactions with the software, its not clear how much opportunity teachers are providing students. To address this challenge, log data for all students in the same class are leveraged together to draw inferences about when class is taking place. First sessions where students in the same class are working simultaneously are identified. Then the earliest transaction of all the students and the latest transaction of all the students in those sessions are assumed to be the start and end time for the class respectively. The total opportunity to work is assumed to be the time between this estimated start and stop. Class sessions were considered to be any of these overlapping work sessions that occurred during the bounds of the school day (between 7am and 3pm).

Once there is some estimate for when each class is taking place, the time that student's work can be contextualized in several ways. Knowing approximately when class begins, it is possible to measure how quickly students start working in each session. This transition speed reflects student's ability to self-regulate when choosing between the types of activities they may be engaged in the start of class and the learning activity. Over the course of the year, an overall start speed ability is calculated as the mean of the observed session start speeds.

While the number of minutes that a student delays beginning work might be one way that student's inform their judgement of the value of getting started at any given moment, they may also make these judgements relative to what their peers are doing. For instance, when most other students are already working, there may be greater incentive to begin work even if the student is delayed just three minutes. Alternatively, if most students are delaying over three minutes, the student may not perceive the disincentive of delaying three minutes to the same degree. Therefore the relative start speed can be calculated by normalizing each students' start speed relative to peers within each session to capture how much faster or slower they are starting relative to the average student.

Within each session, the percentage of session time that students are working can be calculated as the proportion of total time-on-task to session length. This measure captures student's ability to self-regulate within the class session with the pressure from the teacher's monitoring as well as other available distractions. Over the course of the year, an overall ability to self-regulate in class can be calculated as the mean of this per-class percent

session time measure.

Similar to start speed, the percentage of session time might be perceived by students more in relative terms than absolute. A relative session time measure can be calculated by normalizing each session time measure across all students observed within the same session. This relative percent session time measure captures the tendency of students to work longer or shorter than the average student on any given day. Then the overall ability to self-regulate relative to peers is calculated as the mean of this per-class measure. This measure captures the degree to which students are making motivated judgements relative to what they perceive their peers to be doing.

Total overall opportunity can be calculated as the sum of the length of all class sessions to estimate the total expected time students were expected to work. Notably, there are times when students work outside of a class session and the measures discussed thus far do not account for how students may utilize their time outside of class. Percent Opportunity can be calculated as the proportion of total time-on-task total opportunity. This measure aggregates total time worked regardless of whether it was within any particular class session or outside of school, but contextualizes it assuming students are making judgements about time-on-task relative to how much the teacher may have expected them to work. For instance, students may realize they skipped most of a class session on one day because they needed that time to complete an assignment for another class. They then attempt to make up this progress at home later in the day. Percent Opportunity worked will capture the degree that students are choosing to work relative to overall expectation regardless of when they work.

The analysis in this chapter explores the validity of these five measures, start speed, relative start speed, percent session time, relative percent session time, and percent opportunity, relative to the standard time-on-task measure.

5.2.3 Evaluation

5.3 Results

To investigate the value of opportunity information on estimating student self-regulation, this study leverages a correlational analysis to evaluate the construct validity and a regression analysis to evaluate the predictive validity. For the correlation analysis, a pearson correlation between motivational survey measures and three measures, percent opportunity, percent session time, and start speed and a time-on-task control measure. The strength and significance of correlations with each motivational measure is compared relative to the time-on-task measure. To evaluate the predictive validity, each measure is used to predict end of year grade controlling for prior achievement. The model performance is measured using Bayesian information criterion and each measure is compared to both the predictive model including no engagement measure as well as a time-on-task measure.

5.3.1 Class Time Measures

Construct Validity

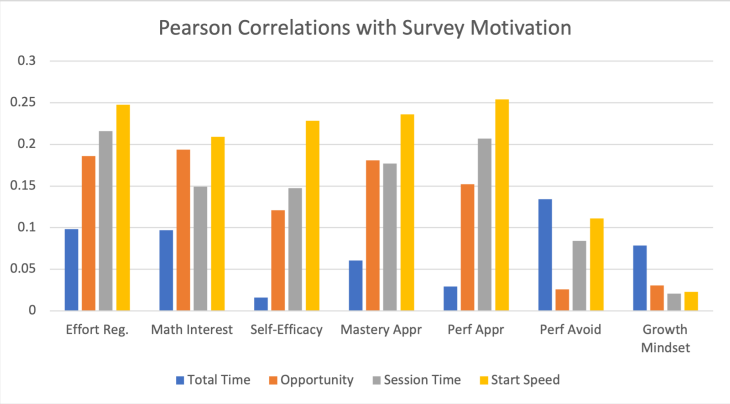

Figure 5.1: Comparing class information measure correlations with motivation survey measures

The correlations between each behavior and motivational measure are shown in Table D.1. The baseline measure, total time, is not significantly correlated with any motivational measure except performance avoidance. As shown in Figure 5.1, every behavioral measure that incorporates class time information is more correlated than total time with most motivational measures. Interestingly, total time has a stronger relationship with performance avoidance and growth mindset than any other measure, though the correlations are only significant for performance avoidance. Further exploratory plots show that where relationships are significant, there is a triangular relationship that indicates each motivational construct is necessary but not sufficient to drive the associated behavior.

Predictive Validity

Table 5.1: Predictive Model Comparison of Class-information measures

Model	BIC
Baseline	476
Total time	459*
Percent Opportunity	450*
Avg Session Length	468*
Start Speed	464*

$*$ - Significant improvement from baseline ($P|t| < 0.05$)

In Table 5.1 below, we see that all engagement measures, including total time, provide explanatory power for end-of-year student achievement in comparison to the baseline model using only prior achievement. Adding class information to total time measures to form the

percent opportunity measure outperforms the total time model. This implies that students that are able to focus in class are achieving better than their peers who are making up for time outside of class controlling for total time worked. However, understanding overall effort is important in predicting overall achievement. This can be seen by Avg Session Length and Start Speed performing significantly worse than total time ($\chi^2 = 7.46$ and $\chi^2 = 6.48$ respectively). A student's start speed limits the amount of the class session that they will be able to work. We see that these two measures are highly correlated ($R = -0.79$), and the models are not significantly different in predicting achievement ($\chi^2 = 0.98$) despite session length explicitly carrying more information about total time worked.

5.3.2 Social Context Measures

Construct Validity

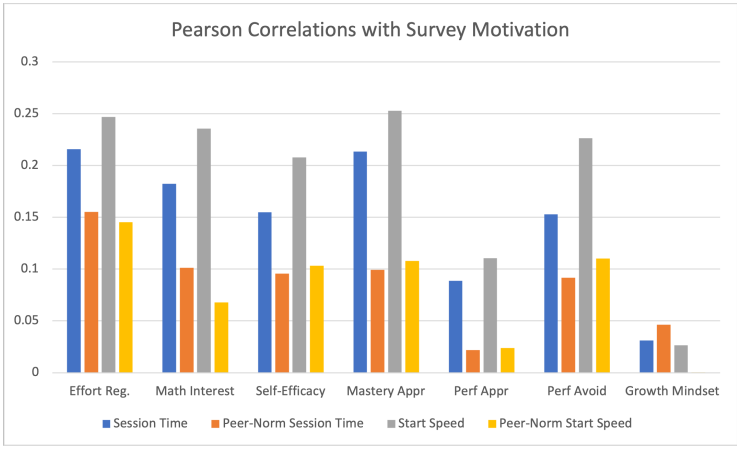

Figure 5.2: comparing Social Information Measure Correlations with Motivation Survey Measures

A comparison of the social and the class information measures are shown in figure 5.2. The specific values shown in this plot along with p-value thresholds can be reviewed in Table D.2 in appendix D. From the comparison shown in figure 5.2, it is evident that the two social measures have consistently weaker correlations than each respective non-social measure. This indicates that reducing start speed or working endurance to the component of these decisions that is driven by peer behaviors does not relate significantly to the motivational measures collected. This is unexpected because performance orientations (approach or avoidance) are defined with respect to how peers are behaving, however, it is unclear whether students perceive engagement time with tutors as a component of performance or if they view success answering problems exclusively as a performance related outcome to optimize. The evidence indicates that overall within this population and context, students did not perceive engagement with the tutor as a measurement outcome on which they might be judged by teachers.

Predictive Validity

As seen in Table 5.2, each of the social information measure models significantly improves upon the baseline models with only prior achievement, there it captures an element of achievement beyond prior knowledge that is reflected in prior achievement. Both social information models outperform their respective class information models. Models knowing only whether students are starting faster or slower than their peers is on par with predicting end of year performance as knowing how much total time students are working over the entire year. Furthermore, measures of whether students worked longer than peers outperforms predictions of end of year grade over total time and even percent opportunity which accounts for time with respect to opportunity. This is especially notable given that this relative session time measure ignores any outside of class work that students may have worked that may contribute to overall knowledge that would translate into differential performance on later academic assessments.

Table 5.2: Predictive Model Comparison of Class-information measures

Model	BIC
Baseline	476
Total time	459*
Avg Session Time	468*
Relative Session Time	440*
Start Speed	464*
Relative Start Speed	457*

$*$ - Significant improvement from baseline ($P|t| < 0.05$)

5.4 Discussion

This evidence taken together provides support for the value of modeling student choices beyond the strictly knowledge-related features that are easily observable through log data. Building measures that leverage some conception of how well students are choosing to utilize the dedicated working time within a class period are shown to capture information about students engagement comparable to if not better than typical total time measures, but also is capturing the influence of various socio-cognitive motivational constructs on student's engagement choices. This evidence indicates that student choices about work take into account the time given and thus a student's ability to self-regulate will better be measured by accounting for how student's are encoding their context to inform their decision making about engagement time with an online tutor.

The evidence for the value of relative social behaviors is more mixed than for classroom time. For both session time and start speed measures, the social-information version of these measures are less correlated with student motivation measures than their respective class information measures. This implies that most student's are not considering what their

making decisions about starting work or continuing to work. This is somewhat counter-intuitive, but it may be possible that the impact of social pressure to begin or stop work is very non-linear, only having a measurable impact when a large majority of students are doing the opposite task. On the other hand, these social measures of engaged behavior are better predictors of end of year grades. However, its unclear whether this predictive power is due to the measure capturing differences in student ability to work with a greater degree of focus, thus leading to more and faster learning or if the measure is capturing some component of how teachers assign grades based on some sense of aggregate engagement. In other words, is some component of teachers' grade assignments attributable to some subjective relative judgement of student classroom conduct?

One of the limitations of this work is that the data was derived from a school implementing a classroom policy where students are expected to engage with the tutor for a full class period. Varying school or classroom policies for when students are expected to work online will have implications for how students perceive the importance of time utilization and therefore how well measures of time-on-task and opportunity utilization will indicate student's ability to self-regulate. Likewise, a combination of student cultural variation and classroom policy may change how students encode the incentives to work in class, introducing measurement confounds that reduce the generalizability of the models developed here. This risk of confounds is inherent to observational data and will require alternative methodologies to reduce these risks.

Chapter 6

Multi-Operational Measurement Models

6.1 Simulated Student Data Evaluation

6.1.1 Overview

The major limitation in the studies in previous chapters is that there is a lack of ground truth on the construct of diligence on which to validate any behavioral measure. Computational simulations are a valuable tool for evaluating methodological questions while eliminating uncertainty about the effect of measurement error on estimation. In the educational data mining community, simulation has been used extensively to explore questions about cognition and learning [31] or tutoring algorithms [71]. However, there is a lack of work that attempts to explore the mechanisms that underlie engagement through simulation. In this chapter, I introduce the development of the Learner Engagement Simulator (LEnS) framework, and then leverage this framework to evaluate the robustness of multi-measure approaches to estimating student self-regulation.

6.1.2 Methods

Design of LEnS Framework

The Learner Engagement Simulator is a python-based real-time simulation framework that enables exploration of real-time fine-grained decision-making models of online learner behavior in context. The framework extends the simulation paradigm found in prior work to enable the emulation of the role of various motivational and self-regulatory constructs on learning and achievement. Prior work focused on simulating the cognition of a learner and/or the policy of the tutor. This approach focused on two main elements in a learning system, the learner and the tutor. Through manipulating either of these two systems, researchers have demonstrated an ability to evaluate the effectiveness of instructional policies, to test theories of human learning, and to support more efficient and scalable authoring of instructional content. This approach to simulating learners assumes learner engagement throughout the simulation, however, engagement is complex and dynamic and a significant explanatory factor in achievement outcomes.

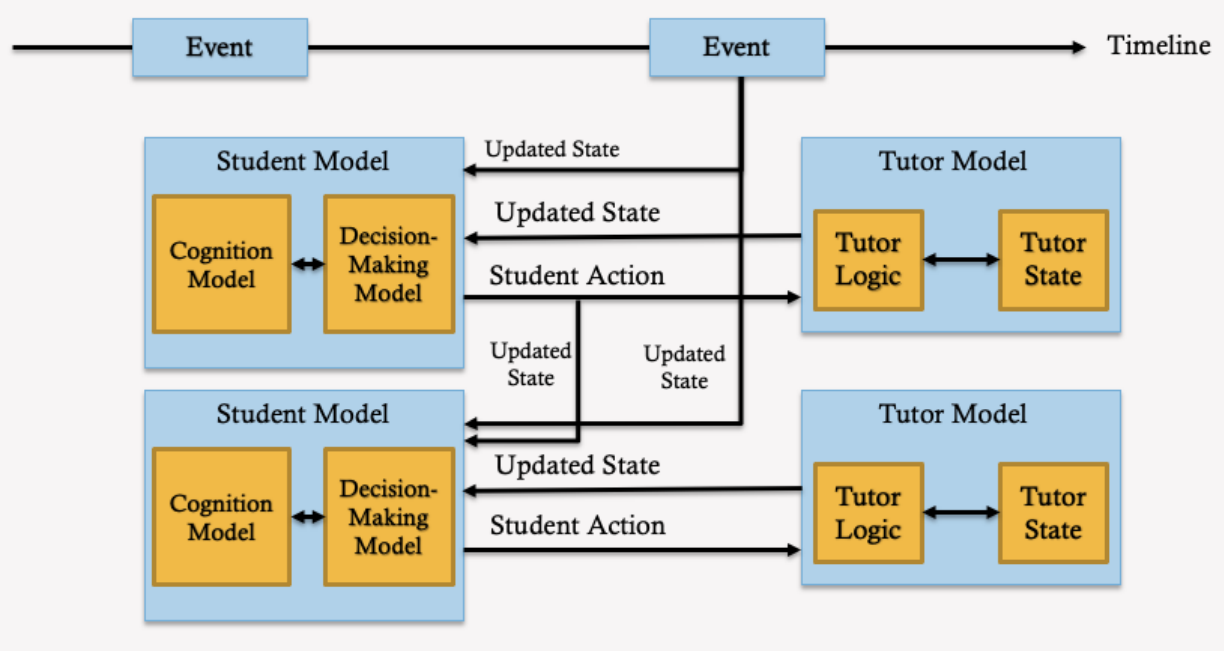

Figure 6.1: Diagram of Learner Engagement Simulation Framework components with messaging interactions

The LEnS framework includes four primary components: the environment model, the tutoring model, the learner cognition model, and the learner decision-making model. Each of these components can be independently defined as appropriate for the research goals for the project. An example of a simple system demonstrating all of the components and their interactions is shown in figure 6.1. As seen in the diagram, simulations defined through this framework can model more complex interactions than just direct learner-to-tutor interaction. Through the environment model, a set of many-to-many interactions between agents can be defined to model more complex social and spatial relationships. In addition to modeling direct agent to agent interaction, the environment model also requires a specification of temporal events. The agent-to-agent and event-to-agent interaction is managed via a publish-subscribe model. These two elements of the environment model, the agent network and the timeline, allow for simulation of a wide range of complex interacting social-technical systems. An example

The timeline forms the foundation of the real-time simulation. The simulation specifies a timeline and a series of events that happen at specific times. With no agent network, the simulation will run without any agent activity. Agents subscribe to the stream of events on the timeline. Each event is represented as a metadata object that encapsulates all relevant information. Agents must have internal logic that filters events for relevant events to respond. For instance, when a class session begins, the event can be represented with a class ID, event type, and a timestamp. Upon receiving the class start event, the learner will begin the process of logging into the tutor.

The agent network in the environment model is implemented using the RabbitMQ

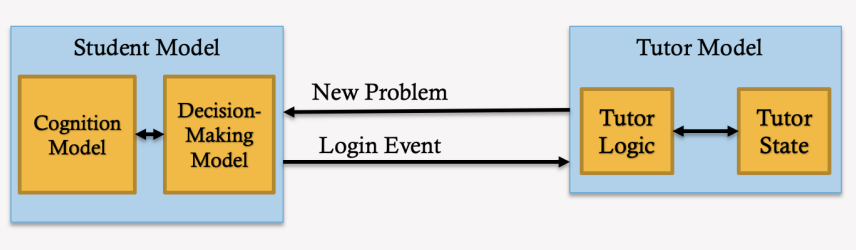

Figure 6.2: Diagram of Learner Engagement Simulation Framework components with messaging interactions

library for agent-to-agent messaging in a publish-subscribe architecture. When agents perform an action, they publish a message that encapsulated metadata about the action that is then published to any subscribers specified in the environment model. An example of a simple tutor-learner event flow is shown in Figure 6.2. In this example, a simple model has a tutor agent subscribed to a learner. When a tutor login attempt event is published, the tutor will process the event, and initialize the tutoring environment and present a problem to begin working. At this point, the tutor will publish an updated state that includes a new problem displayed to work.

In this framework, the tutor model can be as simple as running an actual tutoring software system [74], or emulating a range of problem selection policies [71]. The tutor requires an adapter to translate the input of a student action into the expected input of the underlying tutor model. Likewise, tutor state updates must publish sufficient information that the learner model adapter can determine possible actions and interact.

The learner model consists of two sub-models, the cognition and decision-making model. The learner process begins with a class starting event from the timeline. The decision-making model translates the environment state to determine a set of possible actions. It may call the cognition model to support parameter calculation to support the decision-making process, depending on the actions available. In this example the learner decides to continue their current task of listening to music for 5 minutes until finally logging into the tutor. Once the tutor updates state and presents a problem to be solved, the decision-making model uses the tutor state and environment model to identify possible actions. For actions related to solving the tutor problem, the model uses the cognition model to derive relevant decision parameters. Once all actions are derived and evaluated, the learner uses the decision-making model to make a choice and executes the associated action. If the action is to attempt to solve the problem, then the cognition model is called to produce a response. The result of the action is published by the learner, where the tutor sees the response and processes the input according to its own internal logic.

LEnS introduces innovation in simulations by leveraging a broader context including the social and temporal context that may by influencing learner engagement while they work with tutors. This framework affords the capability to explore hypotheses about how these latent motivational constructs may interact with contextual factors to produce learner behaviors that are observable in traces in log data. Furthermore, empirically validated models can be leveraged to explore the effect of interventions at both the individual and

group level. In the next section, this framework is applied to develop a simulation that emulates the effect of multiple motivational constructs on learner decision-making and how that impacts measures of learner engagement.

Simulator Implementation

This study leverages a simulator developed using the LEnS framework to generate data and evaluate the performance of applying multi-measure behavior scales for estimating latent self-regulation. Interaction between the learner and tutor takes place over the course of multiple class sessions. No social interactions between students or tutor-specific tracking of student state are explicitly modeled, so the simulation focuses on the interaction of motivation and temporal context to influence engagement. The constructs modeled in the student model is varied across simulation runs and the performance of behavioral measures across each set of runs is compared. The following sections will describe the details of the simulation, and the evaluation methodology.

The environment model for the simulation consists of a single learner interacting with a tutor. The timeline schedules a single daily class session. Each simulated student works a total of 20 class sessions, where the session length is random ($\mu = 40$, $\sigma^2 = 8$) with a minimum of 0 and maximum of 60. The variability in session length creates variance in the overall opportunity that students may have to work. Within the available time, the simulation will focus on how students utilize the available time and how this utilization can be leveraged across multiple metrics to estimate the latent self-regulation parameter specified for each student.

The tutor model defines a domain model with 20 curricular units to complete. Each unit has a random number of skills ($\mu = 22$, $\sigma^2 = 23$, $min = 1$) and the skills practiced in each unit are not overlapping with the skills in other units. Each unit consists of a random number of sections ($\mu = 4$, $\sigma^2 = 2$) and the skills within the unit are randomly distributed among the sections. Practice problems are defined as multi-step problems with a random number of steps ($\mu = 10$, $\sigma^2 = 4$), where each step is considered a practice opportunity for a single skill. Problems are generated such that each step is randomly mapped to a skill that is practiced within the section and problems are generated such that there are a maximum of 100 practice opportunities for each skill within the section. The number of problems was selected such that it was extremely unlikely that a skill would require more than the available practice to master. Each skill is generated with 4 parameters common to bayesian knowledge tracing model [6] and two additional parameters to independently model how long it takes for students to solve steps related to this skill. L_0 is the probability that the learner has mastered the skill on the first practice opportunity. $P(T)$ is the probability the learner transitions from unmastered to mastered with each practice. $P(g)$ is the probability a student answers correctly given that they have not mastered the skill. $P(s)$ is the probability a student answers incorrectly given that they have mastered the skill. t_μ and t_{σ^2} are the average and standard deviation respectively of the time to complete a step associated with this skill.

The tutor model traces the progress of each learner on all skills in the domain model using Bayesian knowledge tracing. Given the outcome of the learner's response on the

nth opportunity X_n, the probability that the learner has mastered a the skill after the nth opportunity $P(L_n)$ can be estimated using the equations 6.1-6.3. The tutor uses a mastery threshold of $P(L_n) > 0.9$ to determine when a learner has likely mastered each skill. Problems are selected by first selecting the subset of skills that have not reached mastery in the current section the student is completing. Then a problem is selected at random from the set of available problems that have at least one practice opportunity of the target skill. Due to skill overlap in many of the generated problems, this simple policy likely leads to over-practice of some skills. However, optimal efficiency of learning time is not a goal of this simulation, so this should not impact the study design. For every step of every problem, there are 3 hints available for learners, where the third hint is treated as providing the answer so that the subsequent attempt has 100% chance of correctness.

$$P(L_{n-1}|X_n = 1) = \frac{P(L_{n-1})(1 - P(s))}{P(L_{n-1})(1 - P(s)) + (1 - P(L_{n-1})P(g)} \tag{6.1}$$

$$P(L_{n-1}|X_n = 0) = \frac{P(L_{n-1})P(s)}{P(L_{n-1})P(s) + (1 - P(L_{n-1})(1 - P(g))} \tag{6.2}$$

$$P(L_n|X_n = x_n) = P(L_{n-1}|X_n = x_n) + (1 - P(L_{n-1})|X_n = x_n))P(T) \tag{6.3}$$

The cognition model in the learner model follows the same assumptions as bayesian knowledge tracing. Learners are assumed to have either mastered or not mastered a skill before the first practice opportunity. With each opportunity, there is some chance student's transition from unmastered to mastered. Each learner has a unique cognitive ability parameter, X_{cog}, that represents individual differences in problem solving ability. This parameter is assigned randomly for each learner ($\mu = 0$, $\sigma^2 = 1$). When each learner begins working in a new section, each of the learner's skills, X_{kc}, related to this section are all initialized randomly as either mastered or unmastered. The cognitive ability parameter biases the probability of having initial mastery according to formula EQ 6.4. When attempting a problem, the probability of getting the answer correct are given in equations 6.5 and 6.6.

$$P(X_{kc} = 1) = P(L_0) * (1 + X_{cog}/5) \tag{6.4}$$

$$P(X_n = 1|X_{kc} = 0) = P(g) \tag{6.5}$$

$$P(X_n = 1|X_{kc} = 1) = 1 - P(s) \tag{6.6}$$

The learner's decision model is a stochastic model that operationalizes an expectancy-value paradigm as a mechanism for weighting the likelihood of each choice. Depending on the context, there are a total of six possible actions that learners can choose from: Start working, Attempt Problem, Guess on Problem, Request Hint, Go Off-task, and Stop Working. When learners either have not started working yet or are currently off-task, the only choice is to either stay off-task or to start working. When the learner is currently working on the tutor, the learner can either attempt the problem, guess an answer, request a hint, go off-task, or stop working. Each action has an expected value calculated as the product of the expectancy and total value of the action. The action is chosen randomly with the probability of each action being equal to the action's expected-value as a proportion of

the total of all action expected-values. Once the learner makes a decision, the associated action is executed, and related events are published.

The expectancy for any action was defined as the probability the learner can complete the action successfully. For attempting a problem, the learner will leverage the cognitive model to produce a context specific estimate of answering the question correctly. For this simulation, the cognitive model uses a BKT algorithm to update the learner's internal estimate of skill-specific ability. This model of metacognition and self-efficacy is not expected to be an accurate representation of learner cognition, but cognition with respect to learning is not crucial to this simulation. The key is that learner's have lower self-efficacy on early practice opportunities, and that the learner will experience a wide range of self-efficacies randomly distributed throughout any working session. Guessing an answer is modeled as the type of behavior when learners attempt to answer a question quickly without expending any effort to make an informed guess. The expectancy of guessing is set to 0.05, where it is lower than the average expectancy on all skills at first opportunity, but not so low that the actions is extremely unlikely. Guessing tends to waste time, but when learners guess correctly, no learning occurs as well. All other actions are deterministic, so they have been defined with an expectancy of 1.

The base values for each action were defined with the following behavioral goals. Because the simulation emulates a second-by-second stochastic decision-making process, the starting time of when student begin working are assumed to follow a geometric distribution. When learners have not started working yet, the value of starting was set such that the probability of starting work after 8 minutes would be 5%. Similarly, when students go off-task while working, the value of returning to working was weighted such that the probability of starting work after 4 minutes is 5%. The value of stopping work is a function of the time remaining in the session, where the value of stopping work increases quadratically as the end of the class period approaches. The goal with this is to simulate the mild incentive to finish early while also simulating the strong incentive of learners to stop working by the time class ends. While working, the value of going off-task was tuned such that learners only engaged in this behavior approximately 1-3% of all decisions. Guessing an answer and asking for a hint were set to have very similar values, where asking for a hint is slightly more valuable because it offers the prospect of achievement through perseverance. Attempting to answer the question has a value of 3 times the value of requesting a hint so that learners are expected to answer questions at least 67% of the time.

At baseline, learners across all simulations are modeled with a minimum of two constructs, cognitive ability and diligence. Additionally, self-efficacy and intrinsic interest in the domain are defined in the decision-making models using additional parameters. There are four sets of students generated for comparison of behaviors. The baseline students vary only in terms of cognitive ability and diligence parameters. There is one set of students that also vary in terms their self-efficacy parameter. There is a third set that vary in terms of their intrinsic interest, and a fourth set varies in terms of self-efficacy and intrinsic interest.

As described previously, cognitive ability is a parameter that influences the operation of the cognitive model. Diligence is a core parameter in the function of the decision-making model. Within a particular context, each action is considered either a diligent or a non-diligent action. More diligent students are simulated to be more able to prioritize diligence

actions. This is implemented by weighting the expected-value of diligent actions proportionally to the learner's diligence parameter. The result is that more diligent students will tend to engage in any diligent action more often than a less diligent student in the same context. When starting class, the impact of diligence on re-weighting the value of starting was derived such that a student that is one standard deviation over the mean diligence level would have a 5% chance to start after 5 minutes as opposed to 8 minutes.

Intrinsic interest and self-efficacy are implemented similarly by weighting the expectancy and values for specific actions. Students with greater levels of Intrinsic interest will have higher value associated with attempting a problem and requesting a hint because both actions are associated with engaging cognitively with an activity that is intrinsically rewarding. Likewise, more intrinsically interested students will see less value in guessing an answer. When either deciding to start class or resume working when off-task, the effect of intrinsic interest is equivalent to the effect of diligence on increasing the learner's average start speed. In this way, learners with greater intrinsic interest are expected to start faster, go off-task less, work longer, and learn more. Self-efficacy is modeled at the domain level, where learners have some more domain general belief in their ability to solve problems which biases their own expectancy when considering a problem. This is implemented by having learners, with self-efficacy at one standard deviation greater than the mean, will have an expectancy that is up to 1.2 times greater than an average learner. The result is that lower self-efficacy learners will tend to spend more time requesting hints, guessing, going off-task, and even stopping earlier.

This set of simulation models approximates the complexities of confounding constructs on motivated decision-making. In particular, decisions specifically related to time-on-task and learning efficiency are multiply confounded and not uniquely identifiable using strictly log data. This is a major hurdle in estimating student self-regulation through observable behaviors, and this multiple confounding is manipulated in the following simulation study to demonstrate the benefits of multi-measure instruments on the task of estimation with observational behavior data.

Evaluation

In this study, four simulation runs are compared. Each run consists of 100 learners where the learner models vary across runs based on the latent constructs they implement. The baseline simulation includes learner models that implement only cognitive ability and diligence. Three comparison simulations are also run each implementing confounding constructs in addition to the ones implemented in the baseline condition. The first implements self-efficacy. The second implements intrinsic interest. The third is a combination simulation that implements both self-efficacy and intrinsic interest.

Within each simulation run, each of the learner parameters are generated independently and randomly ($\mu = 0$, $\sigma^2 = 1$). Each simulated student works a total of 20 class sessions, where the session length is random ($\mu = 40$, $\sigma^2 = 8$) with a bounded range of zero to sixty minutes. For each simulation, learner decisions are logged, and interactions with the tutor are logged in a format compatible with the Datashop learner transaction data standard [35].

The performance of four behavioral measures will be compared. The percent opportunity measure, X_{opp}, is calculated as the proportion of the total duration of all diligent transactions (ie: time-on-task) to the total of all session lengths(EQ 6.7). The percent off-task measure, X_{ot}, is how likely the learner is to go off-task and is calculated as the proportion of all transactions that are off-task to the total number of transactions (EQ 6.8). The start speed measure, X_{start}, is the mean time it takes for learners to start working and is calculated as the average of the difference between the first action and the start of session(EQ 6.9). The early finish measure, X_{finish}, is the average time before the end of class that learners stop working and is calculated as the mean of the difference between the last action and the end of the session (EQ 6.10).

$$X_{opp} = \frac{\sum_{i=transactions} t_{i,duration} * I_{i,diligence}}{\sum_{i=1}^{x_{SessionCount}} t_{sessionlength}} \tag{6.7}$$

$$X_{ot} = \frac{x_{OfftaskCount}}{x_{TransactionCount}} \tag{6.8}$$

$$X_{start} = \frac{1}{x_{SessionCount}} * \sum_{i=1}^{x_{SessionCount}} x_{i,StartSpeed} \tag{6.9}$$

$$X_{finish} = \frac{1}{x_{SessionCount}} * \sum_{i=1}^{x_{SessionCount}} x_{i,EarlyFinish} \tag{6.10}$$

In addition to the individual behavioral measures, a set of composite measures will be calculated using every combinations of 2,3, and 4 behavioral measures. In each composite measure, a latent factor model is calculated by performing a factor analysis with 1 latent factor. The data is projected onto the latent factor using the factor loadings to calculate the composite measure. Overall, 15 diligence estimators will be compared for each simulation run. For each estimator, the square of the Pearson correlation coefficient (r^2) is calculated between the estimator and the learner diligence parameters for each simulation run.

6.1.3 Results

To start, we validated the effect of the experimental manipulation of the effect of confounding factors on recovering a measure of the diligence parameter. Figure 6.3 shows performance of each of the individual behavior measures on its ability to recover the latent diligence parameter as measured by r^2. The effect of the experimental manipulation is evident in the general downward trend of the r^2 values as each confound is introduced to the simulation. Relative to baseline, the introduction of confounds tends to reduce correlations with particular behaviors, and the combined simulation run shows further reduced correlations. In the baseline condition, the estimation challenge with these simulated parameters is evident in the moderate to low correlations with r^2 ranging from 0.34 to 0.5. With only 20 sessions, as in typical real-life applications, there is limited data to estimate parameters for each behavior given the variance of the underlying generating function. If evaluating the performance of these indicators on student data, opportunity would have

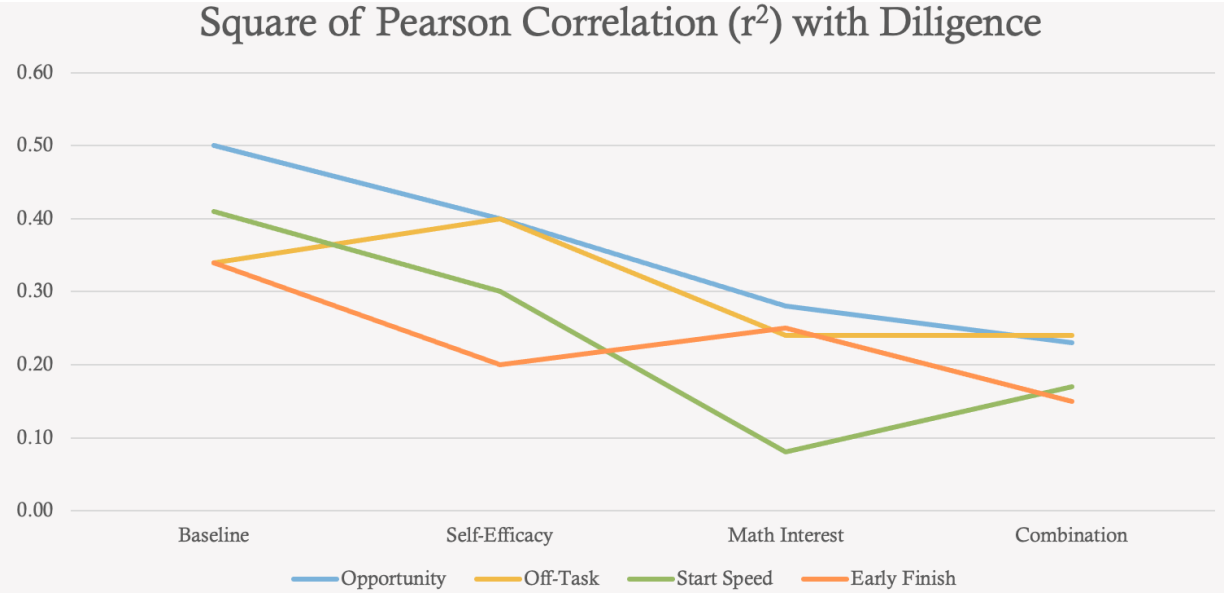

Figure 6.3: Baseline measure: r^2 comparing individual behavior measures with latent diligence parameter on simulated student data

been selected as the best performing measure with the exception of the combined confounds condition where Off-Task, $r^2 = 0.24$, slightly outperformed opportunity, $r^2 = 0.23$, though the difference was very small.

To compare the benefit of leveraging multiple operationalizations in the instrument, the best performing combination of measures is compared for instruments of overall size ranging from 1 to 4, where size=1 is the set of baseline measures discussed previously. Figure 6.4 demonstrates the main effect of using multi-measure estimators as a general trend of improving performance over baseline within each simulation condition. This is again evident in the within condition comparison of measures. However, it was unexpected that the two-measure instruments outperformed three and four-measure instruments in all but the baseline condition. It was expected that there would be a general benefit for including more measurements. However, a closer analysis of the results for each instrument, which can be referenced in appendix D.2.1, is that the combination of opportunity and off-task measurements performed the best. Given that opportunity is derivative of start speed and early finishing, while off-task is independent of both of the remaining three measures, it appears that the in this simulation, the extra information offered by delineating start and finish behaviors is potentially overwhelmed by the noise that such measures introduced. This implies that some of the results presented here require further theoretical and experimental exploration to understand the relationship between the signal-to-noise ratio of individual measures and the measurement power they offer to the overall instrument.

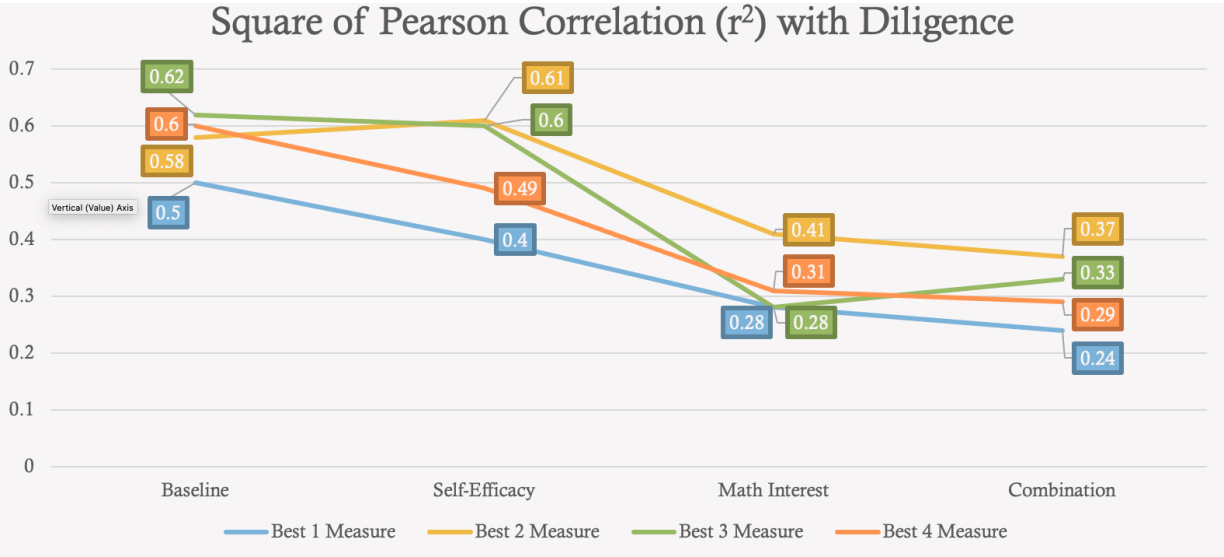

Figure 6.4: Comparison to baseline of r^2 of multi-operational instrument measures of varying size with latent diligence parameter on simulated student data

6.1.4 Discussion

In this study, the value of leveraging multiple behavioral measures to estimate learner's latent ability to self-regulate is demonstrated through manipulation of confounding constructs in simulated data. In general, self-regulation is a challenging construct to measure with observational data because the construct is impossible to identify given the limitations of the data. Nonetheless, educational technology regularly leverage engagement analytics as an indicator of students' diligence, and therefore there is a benefit to improving the accuracy of these metrics. This simulation study demonstrates that leveraging a multi-measure instrument approach to estimation will offer an improvement over contemporary single measure approaches.

One challenge with this approach is evident in how the instruments with greater than 2 measures could be improved by possibly introducing more independent measures of diligence. This is due to the redundant information that is contained in the percent opportunity measure relative to the start and finish measures. This highlights the continued need for intelligent design of behavioral metrics that capture student's self-regulation, however, it also demonstrates that one new consideration in addressing mono-operational bias is in empirically validating the assumptions that new individual behavioral measures provide a marginal benefit to an existing instrument. This is a common practice in the development of survey-based psychometric instruments, and research and methodologies in this area can be leveraged to tackle similar challenges in developing multi-measure behavioral instruments.

In this study, a simple approach of using estimating a latent common factor across multiple aggregate behavioral measures was leveraged as a methodology for addressing mono-operation bias. The key finding is that different behaviors implicate confounds to

self-regulation differently and that operationalizing the target construct in multiple ways can mitigate measurement biases. The natural variation in the underlying structural model of how self-regulation and its confounds are implicated in each behavior can be leveraged as additional information to improve on model estimation. Therefore, a natural next step in this work will be to explore more advanced modeling approaches to leveraging data across multiple behaviors, incorporating rich contextual information, to build more robust estimation models.

6.2 Real Student Data Evaluation

6.2.1 Overview

One of the great challenges with observational data is attempting to attribute variance to signal or noise. Studies in prior chapters have demonstrated that student diligence manifests in behaviors observable through fine-grained learning log data. Furthermore, some motivational confounds can be accounted for through intelligent feature engineering. However, the models developed thus far show either strong construct validity or predictive validity, but not both. Also, the measures tend to be unreliable and unstable, so they are difficult to use as a psychometric measure. The challenge with observational data arises from the fact that a particular behavior is driven by some set factors, and this structure can be elicited in the data. However, not all of the factors can be observed in the data to a degree that would enable identifying the signal from the noise. With motivation this challenge is compounded because there is always the possibility that unexpected events like a conflict with a close friend can drive a change in behavior. Because these events are not easily encoded in the data, the behavioral measure is prone to these confounds, thus making the measure less reliable.

One approach to address this risk is to diversify the measures used to estimate student diligence. While a particular behavior, such as choosing how quickly to start class, will be driven by a set of factors including the target measure with some predictable structure, a different behavior may also implicate the target measure while having a different relational structure to confounding factors. By leveraging multiple behaviors that operationalize diligence differently, it is now possible to estimate diligence as a latent common factor across the set of behaviors. The risk of any confounding source is reduced by the fact that a sufficiently diverse enough collection of measures will contain some measures that are not influenced by a particular confounding source, and therefore they will not reflect the variation driven by this source of noise. In this chapter, I attempt to explore the viability of this multi-operationalization approach to estimating diligence with log-based behavior data.

6.2.2 Methods

This study compares the psychometric attributes of a combined latent measure to the multiple operationalizations of diligence developed in prior chapters. The same dataset

that was used in prior chapters is used again for this study [47]. The multi-operational measure is defined as the common latent factor across the set of defined diligence measures. Horn's Parallel analysis [1] is applied to determine the appropriate number of latent factors that best fits the data as described by the full set of individual diligence measures. Using the resulting factor loadings, the data is projected in the latent factor space and the psychometric attributes of the resulting latent factor measures are evaluated relative to the individual measures.

The diligence measures used are attendance, percent opportunity, start speed, relative start speed, early finish, relative early finish, percent session time, relative percent session time, and gaming tendency. Attendance is calculated as the proportion of all inferred class sessions where the student is observed working with at least one transaction. Percent opportunity is calculated as the ratio of the total time working to the total opportunity provided for the class, where the total opportunity is the sum of the length of all class sessions for a class. Because students can work outside of class, this measure can be greater than one. Start speed is the average delay time between the start of a session and the first action of the student for all sessions where the student is observed working. Relative start speed is an average over all observed working sessions of the student's start time normalized across the set of observed start times for that class session. Early finish is the average over all observed working sessions of the time between the last time the student is observed working and the end of the session. The relative early finish is an average over all observed working sessions of the student's early finish time normalized across the set of observed early finish times for that class session. The percent session time is the average of the proportion of each class session that the student is observed working. The relative percent session time is the average over all observed working sessions for the student of the percent session time normalized across the set of percent session time measures for all students observed for that session. The gaming tendency is the probability the student may engage in a gaming behavior, controlling for the probability of gaming due to working on problems from a particular section of the curriculum. Gaming is defined according the the expert defined heuristic model defined in [58]. The tendency is estimated using a random effects model predicting the probability of gaming on a particular problem-step, with random intercepts for each student and section of the curriculum. The gaming tendency is defined as the fitted random intercept for each student.

Once the latent factors have been calculated, the construct validity, predictive validity, reliability, and stability of the latent factors will be compared to the individual measures to compare the relative performance of the combined factor. To evaluate the construct validity, the correlations of the latent factors with the survey-based motivational measures is compared to correlations of the individual measure with the survey measures. To evaluate the predictive validity, each of the individual measures and the latent factors are added to a baseline random effects regression model predicting end-of-year grade given prior achievement with a random intercept for class membership. The Bayesian Information Criterion (BIC) is calculated to compare the predictive performance of each model.

In order to calculate the reliability and stability of the measures, the data is split into halves. To calculate stability, the data is split by date, where the first half includes all data that occurs before the end of December. The second half is all data that occurs after

last day of December. To calculate the reliability, the school year is divided into 36 weeks and the weeks are randomly divided into two equal sets of 18. The data is then divided according to which set of weeks the data belongs. Once the data has been divided, each of the individual measures is calculated for each half. The latent factor analysis is performed on the first half of the data to calculate the factor loadings. Then the loadings are used to project the data in each half respectively to calculate the latent factor projections. For each measure, the stability and reliability are computed as the pearson correlation between the values calculated from each half of the data. In this dataset, one of the 226 students was found to have data in only the first half of the year, so that student was removed from the stability analysis.

6.2.3 Analysis

Factor Analysis

Applying Horn's parallel analysis, 3 latent factors emerge as the best fit number of latent factors for this dataset. The loadings for each latent factor are shown in table 6.1. Factor 0 could best be interpreted as the diligence factor for which the instrument was designed. This factor captures the degree that students are starting quickly, before peers, ending later, and working longer. Notably, gaming does not load strongly on any of the factors. This is somewhat surprising because rationally gaming should be an additional application of self-regulation and prior studies reinforced this hypothesis. However, decisions about gaming may be more influenced by other factors related to the problem being solved than by the dominant factors influencing time-management choices. Factor 1, the procrastinator factor, captures the extent that students attend class but don't utilize as much of the opportunity given to them, predominantly because they tend to start late but then work most of the remaining time. Factor 2, the slacker factor, captures the degree to which students are avoiding work by stopping early. It is primarily defined by finishing early and working less session time. As hypothesized, the dominant factor is a diligence-like factor, though some measures, such as gaming and early finishing, loaded very lightly on this measure.

Table 6.1: Factor Loadings

	F0 (Diligence)	F1 (Procrastinator)	F2 (Slacker)
Opportunity	0.831	0.327	0.220
Session Time	0.874	0.050	-0.194
Peer Norm Session Time	0.811	-0.098	-0.100
Start Speed	-0.890	0.391	-0.001
Peer Norm Start Speed	-0.842	0.514	0.037
Early Finish	-0.386	-0.530	0.661
Peer Norm Early Finish	-0.044	-0.559	0.209
Attendance	0.453	0.623	0.638
Gaming	-0.163	-0.209	-0.070

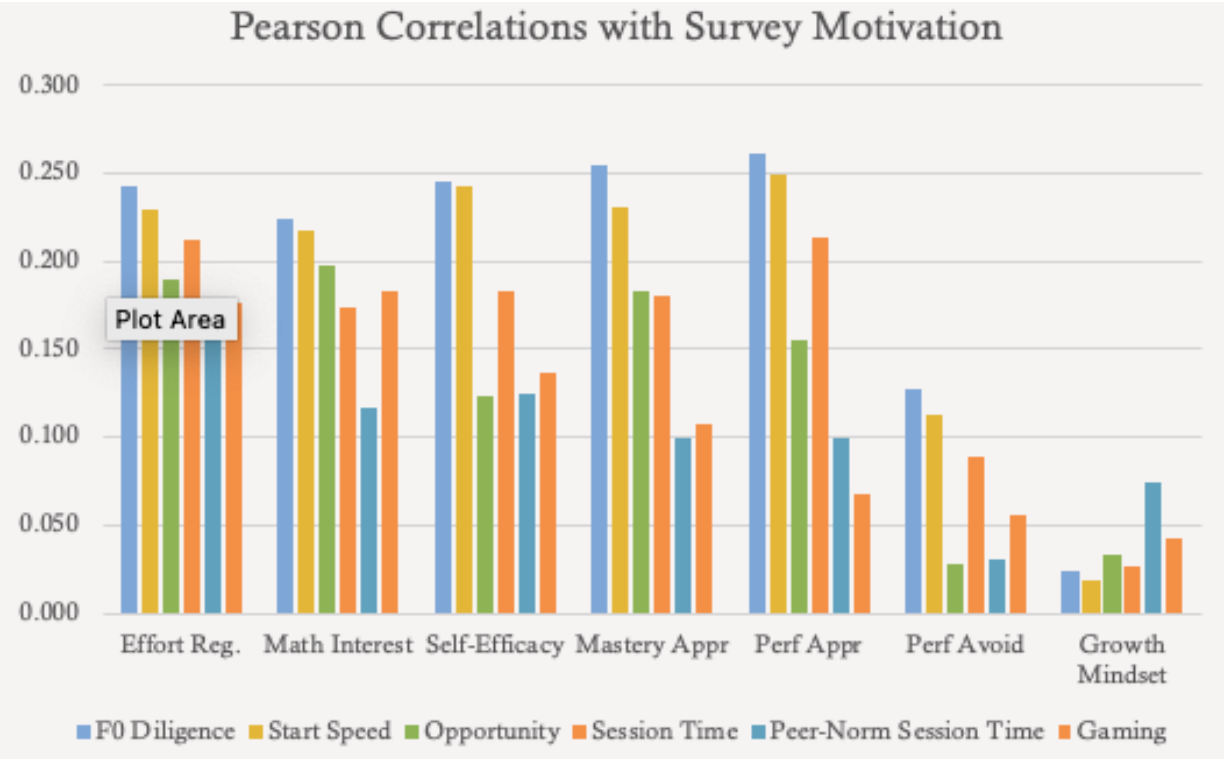

Figure 6.5: Comparing correlations of multi-measure diligence factor with motivational survey measures to correlations of raw behavioral measures of the same survey measure

Construct Validity

The detailed tables of all correlation results can be reviewed in appendix D. Inspection of the direction of the correlations of each factor with achievement measures supports the interpretations of each factor. The diligence factor shows a positive relationship with all achievement measures. Likewise, the procrastination and slacker factors have a negative relationship with each achievement measure as expected. A graphical summary of the correlations of the motivational survey measures with each of the raw behavior measures as well as the diligence factor can be seen in Figure 6.5. As seen in the analysis in chapter 5, the start speed measures is still the raw behavioral measure that best indicates student motivation across the board. The only exception is with growth mindset. Relative Early finishing is the best indicator of growth mindset where students that work until closer to the end of class are more associated with possessing a more growth mindset. No other behavioral measure including the latent factors were significantly related to growth mindset. In comparison to the best indicators of motivation, the diligence factor shows the strongest correlation with each motivational measure in comparison to any of the individual behavioral measures with the exception of growth mindset. Interestingly, the diligence factor appears to capture the similar facets of self-regulation that are indicated by start speed as it shows very similar correlations with motivation measures with a small but consistently stronger correlation for each measure.

Predictive Validity

Any measure of self-regulation should explain achievement above and beyond prior knowledge. As seen in Table 6.2, the procrastination and slacker factors are not significant in each respective regression (p=0.2 and 0.17 respectively), and each model fails to improve over the baseline using only prior achievement (BIC = 456 and 460 respectively in comparison to 452). However, the diligence factor does significantly (p=0.002) predict end-of-year achievement accounting for prior achievement as expected. In fact, the diligence measure performs similarly (BIC=430) to the best predictor, opportunity (BIC=428). A-priori, it is not clear whether diligence should better explain achievement over more direct measures of effort such as percent opportunity or relative session time, which capture measures of either overall work or work in class relative to peers. Class grades are a composite measure of performance on various assignments as well as other measures such as attendance and class conduct, the influence of diligence on achievement is more indirect than some of these other more predictive measures, which likely explains the weaker but still significant predictive results.

Table 6.2: Comparing multi-measure latent factors performance predicting end-of-year grade to that of individual behavior measures

Model	BIC
Baseline	453
F0 (Diligence)	430*
F1 (Procrastinator)	443
F2 (Slacker)	447
Attendance	443*
Pct Opportunity	428*
Gaming	450*
Start Speed	445*
Relative Start Speed	444*
Early Finish	461
Relative Early Finish	455*
Pct Session	451*
Relative Pct Session	431*

$*$ - Significant improvement from baseline ($p < 0.05$)

Stability And Reliability

The stability and reliability of each individual measure and latent factor are shown in figures 6.7 and 6.6 respectively. The most stable individual measure is session time (R=0.67) with each of the start speed, session time, and opportunity measures all having a stability between 0.55 and 0.67. None of the individual measures pass the heuristic threshold of 0.7 for an acceptably stable measure [5]. The same set of measures are also the most reliable measures with R ranging from 0.63 to 0.79 with four out of five measures having

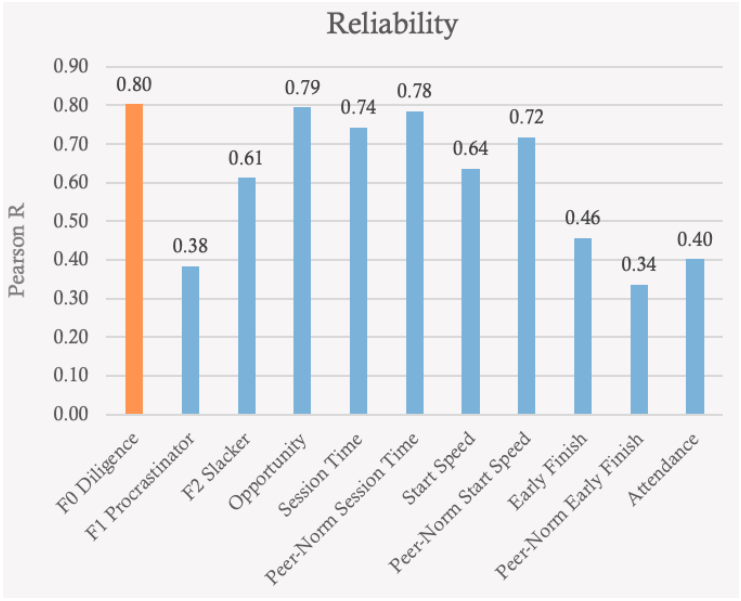

Figure 6.6: Comparing reliability of multi-measure diligence factor to raw behavioral measures

an acceptable reliability of $|R| > 0.7$. The diligence factor is acceptably stable and reliable, and it is more stable (R=0.71) and reliable (R=0.80) than the most stable or individual measures, raw session time and percent opportunity respectively.

6.2.4 Discussion

One surprising result from inspecting the factor loadings is that gaming behaviors did not align closely with these other time-on-task related behaviors. When developing psychometric instruments, the low agreement between gaming and other measures implies that gaming should be dropped to improve the instrument. However, in this set of measures, gaming is the only fine-grained behavioral measure included in the instrument. The overall instrument analyzed here consists of measures that are focused on students' abilities to self-regulate with respect to their local environment. Though there is a non-trivial number of measures in the instrument, most of them may share some set of confounding factors that are not influential during the moment-by-moment learning process. This implies that the gaming measure is possibly in fact introducing desirable diversity to the instrument. This is a question that is worth exploring in future studies.

A summary of the results of the psychometric evaluation of the multi-operationalized diligence instrument relative to individual behavioral measures is shown in Figure 6.3. overall, the evidence supports the conclusion that a relatively simple multi-operationalized instrument approach yields an overall improved psychometric measure relative to any of the proposed individual measures. The diligence factor shows a significant relationship with almost all motivational measures, with a mild improvement over the best indicator

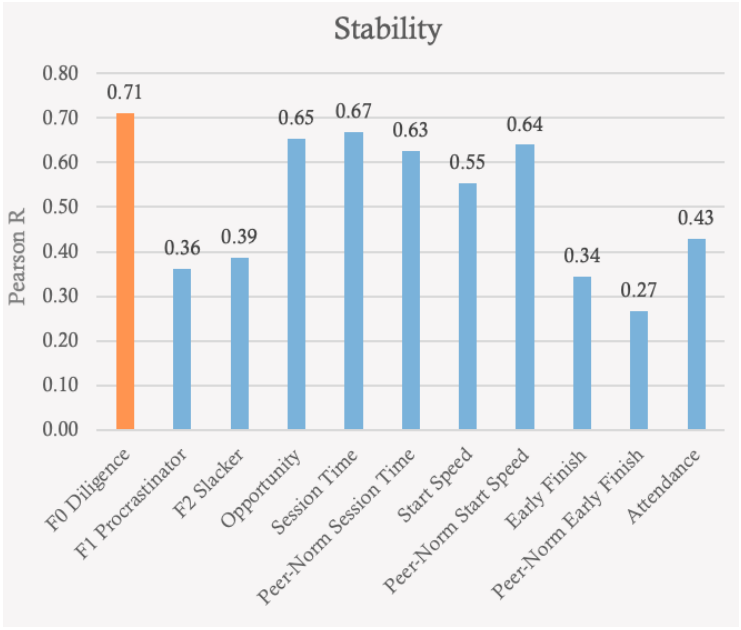

Figure 6.7: Comparing stability of multi-measure diligence factor to raw behavioral measures

Table 6.3: Summary of comparison of psychometric characteristics of multi-operational diligence instrument

	Best Individual Measure	Relative Diligence Performance
Correlation	Start Speed	Better
Prediction	Opportunity	Comparable
Stability	Session Time	Better
Reliability	Opportunity	Comparable

of motivation raw start speed. However, raw start speed is only moderately stable and reliable. Opportunity and session time are much more stable and reliable measures than raw start speed, but they trade reliability and stability for construct validity. The combined diligence factor appears to not require these trade-offs, demonstrating superior construct validity, stability, and reliability.

One of the limitations in this study is that the latent measure captured by this behavioral measure scale is not validated against an independent measure of student diligence. This dataset lacks a more direct measure of student diligence such as one collected by the Academic Diligence Task [55]. Furthermore, evidence for predictive validity is supportive but weaker than results from construct validity, stability, and reliability analyses. While student ability to self-regulate will influence many behaviors such as time-on-task and degree of cognitive engagement, other factors such as student knowledge and student-teacher relationship may also have direct influences on outcomes that influence end-of-year grade. Thus, behavioral measures that better capture the influences of these other factors may better predict achievement than diligence.

Chapter 7

Discussion

7.1 Overview

In summary, this dissertation addresses the challenge of using log data to identify a reliable measure of the individual differences between students capacity to engage in learning. Over the course of five studies, I explored evidence for motivational factors in log data that can be accounted for to more accurately identify diligence. Then I utilized these factors to demonstrate, with simulated and real student data, that leveraging multiple operationalizations reduces bias and identifies a reliable and predictive measure of diligence. A reliable measure of diligence enables new forms of adaptation and student support. In this chapter, I explore the implications of this contribution for researchers and designers of educational applications.

7.2 Adapting instruction to optimize engagement and learning

Skilled human tutors not only respond to the knowledge specific needs of learners, but also adapt to the unique personality of the learner and signals of their motivation throughout the day. Prior researchers have focused on gauging the knowledge and skills of learners [44] as well as the moment-by-moment fluctuations in motivation [33], and using this information to inform problem or activity selection. A reliable measure of diligence enables problem and activity selection that is more personalized to how likely the next activity may disengage the student.

Applications that encode student diligence can use that information to select problems and activities that support engagement for less diligent students. One of the challenges with instruction is that not everything that needs to be learned can be easy and pleasant. There are inherently more and less challenging and engaging materials in a curriculum. Learners must work through all the material. More diligent students will more readily engage in a learning task regardless of its characteristics, while less diligent students are more likely to be slow to engage and also to disengage early without utilizing most of the opportunity afforded them. Diligence measures can provide an ability for applications to extend prior work on predictors of disengagement or gaming [82, 57] by personalizing to

students' diligence. These predictions can be balanced in turn against the various learning parameters that might be weighted for problem or activity selection.

Attempting to push certain practice problems to another time creates a need to plan the timing of certain practice and activities both within and across sessions. Not all skills and knowledge are best suited for the same types of practice [45]. Some factual knowledge acquisition and procedural fluency development are more suitable for adaptation to more engaging or activities with lower barriers to starting. Nonetheless, every curriculum includes some significant quantity of materials that is more disengaging, but nonetheless must be completed. Applications can attempt to optimize learning time by mitigating this disengagement risk or time the riskier practice to maximize some balance between overall engagement time completing more disengaging material. One example might be to consider spaced practice needs by pushing more fluency building exercises to sessions long after initially learning some new material. This allows one to focus on more challenging and likely disengaging materials in the later part of sessions, while allowing more fluency building early in a session to reduce disengagement risk and maximize overall engagement time.

While diligence measurement opens new avenues for personalizing learning experiences, it also exposes a new space of questions for researchers to explore. Leveraging predictive models of quitting and disengagement in real-time to anticipate and intervene on learner motivation is still a very under-explored area. The variability in motivation that influences engagement decisions as students work through problems of varying difficulty and experience strings of successes and failures is still an under-modeled area of research. The complexity of modeling dynamics in these multi-variate time-series is difficult even with the scale of student data available. However, the LEnS framework can allow researchers to explore hypotheses and narrow the space of possibilities before testing more specific models on student data. Additionally, this framework allows for evaluating the longitudinal effects of any tutors new problem selection algorithm on overall student learning.

7.3 Enabling motivation interventions with multi-operational models

Knowing the individual differences in student's diligence is valuable as teachers are becoming acquainted with their students. However, as the school-year progresses, teachers monitor students for changes in behavior that might point to an issue that needs to be addressed. Leveraging multiple-operationalizations enables more reliable measurement of student characteristics, however, variability across operationalizations can carry diagnostic information about student's motivations. This variation can be leveraged as an asset instead of just bias information to filter from the measurement model.

Individualizing instruction is more than just accelerating struggling students, but also supporting all students in learning to the best of their abilities. For instance, the second latent factor that emerged from the analysis in chapter 6 was a procrastination factor, characterized by a slow start speed and moderate to high diligence values along most other measures. This implies there might be a student that is quite diligent but has some strong

and consistent distractions at the start of class such as always walking into class while talking with a good friend from the previous class. This inevitably leads to poor starting speeds because of the strong temptation to continue the conversation. Once this student starts, they work hard on all the problems and they utilize the rest of the remaining opportunity available. In this case, the majority of the behavioral measures indicate a moderate to highly diligent student, though the student will appear low diligence according to the start speed measure. This disagreement in measures can be indicative of a motivational confound, the distracting friend. This can be informative for educators to either pay special attention to the student near the start of class or talk to the student directly to identify whether there was any way to improve their classroom diligence. Providing analytic support and diagnostic scaffolds within educator dashboards can be an important application of the diligence instrument.

Start speed in particular was an interesting measure that emerged from the analysis in chapter 5 because it was the strongest indicator of student motivation. One possible reason for this is likely related to the fact that start speed decisions are not influenced by the specific attributes of the learning activity not under the students control such as problem difficulty because the student has not begun working. Instead, the student balances domain and tutor-specific perceptions of value against their current activity or other alternatives. By removing variation in observed behavior due to interactions with the tutoring algorithm, start speed behaviors can be a more straightforward indicator of motivation. Likewise, because this behavior is largely influenced by factors more directly under the student's control as opposed to the tutoring application, it can be more appealing as a target for behavioral change. Because starting slowly necessarily constrains the potential instructional time available to a student, improving this behavior also opens the door for more overall engagement. For these reasons, applications should consider specifically incorporating interventions and analytic support for start speed improvements either mediated by teachers or directly to students through some form of open learner model.

It may also be valuable to use multiple behavioral measures to gauge whether a student on a particular day might appear to be experiencing some extraordinary motivational impacts. One of the challenges with individual measures is that variability over time is expected, so it is hard to gauge whether an outlier value is indicative of some motivational impact that is worth further attention. However, with multiple behavioral indicators, it may be possible to have more confidence when seeing a consistent trend across multiple behavioral indicators for the day. For instance, a student starting late one day, may not be noteworthy. However, if they are also showing a lot of gaming and off-task behaviors, and then they stop working early, educators may more confidently act on this scenario if dashboards are able to surface this sort of behavioral anomaly. In fact, this paradigm of intervention support can help teachers identify students that might be feeling particularly disengaged as well as those students who are demonstrating exceptional engagement. Enabling more immediate positive reinforcement can help students more easily identify what practices they are doing that are successful. Providing analytic support for both positive and negative motivational support can be a powerful tool for teachers to guide their students to greater engagement and higher achievement.

Chapter 8

Conclusion

Effective instruction not only addresses the learning needs of students, but also recognizes their unique motivations and adapts instruct to support both of these dimensions. This dissertation advances the capacity of technology enhanced learning environments to recognize the motivational dimension of students. To this end, this dissertation makes several contributions.

1. In the learning sciences, this work contributes to the understanding of how student's capacity to self-regulate interacts with the cognitive, temporal, and social contexts during classroom learning to form patterns of engagement.
2. To the learning analytics community, this work contributes a fine-grained, log-data based model of student diligence that can be leveraged to assess and support students' engagement.
3. For the educational data mining community, this work demonstrates the value of leveraging multiple-operationalizations when building behavior-based measurement models of motivation.
4. For the human-computer interaction community, this work develops the LEnS Framework for furthering the use of simulations to model how motivational constructs drive the behavior of systems of learners and educational technology within a learning environment.

I invite researchers and developers to build on this work to explore how motivationally-aware applications might improve learner experiences or better support educators in connecting with their students.

Appendix A

Appendix A: Dataset Information

A.1 Demographic Information

In the following tables, the counts for each demographic category are shown for three sets of data. The "All" column describes the full set of 271 students with demographic information collected. The "No TX" column describes the set of 11 students with demogrphic information but no observed tutor transactions. The "Complete" column describes the set of 226 students that have observed tutor transactions, demographic information, and a complete set of achievement and motivational survey measures at the start of the year.

Table A.1: Sample size for ethnicity over each subset

	All	Incomplete	Complete
White	265	42	223
Non-white	6	3	3

Table A.2: Sample size for gender over each subset

	All	Incomplete	Complete
Male	131	24	107
Female	139	20	118

Table A.3: Sample size for Free/Reduced Lunch Status over each subset

	All	Incomplete	Complete
Not F/R	210	33	177
Free/Reduced Lunch	61	12	49

Table A.4: Sample size for Special Education Status over each subset

	All	Incomplete	Complete
Not Special Ed	233	39	194
Special Ed	38	6	32
Not Gifted	269	45	224
Gifted	2	0	2

Table A.5: Observed Gaming Frequency by Section including over non-hard and hard section subsets

	Count	Mean Gaming	Stdev Gaming	Min Gaming	Max Gaming	Median Gaming
All Sections	237	1.95%	1.7%	0	32%	
All Sections eee-without low obs	196	2.48%	3.64%	0	28.3%	1.6%
Non-Hard Sections	156	1.40%	0.83%	0	3.23%	1.24%
Hard Sections	40	6.70%	6.37%	3.25	28.3%	4.27%

Appendix B

Appendix B: Gaming Detector

This dissertation leverages the model of gaming developed by [58] to annotate transactions as gamed. This model identifies a set of patterns of transactions that experts identify as gamed patterns. A student is determined to be gaming at some time if a series of transactions matches an identified transaction. For instance, a common pattern is when students enter the same or a very similar answer into multiple places without answering correctly, effectively guessing where a calculation result belongs without understanding the organization of the problem. Another common pattern is when students ask for help without taking much time to consider the problem, followed shortly after by an incorrect input. In this case, the student appears to be using the help facility to get an answer but is not taking enough time to use the information provided to derive an answer.

To apply the gaming detector from [58], first individual transactions are labelled using a set of heuristic threshold given in B.1. Then the transactions are reviewed sequentially to identify any matches to patterns shown in figure B.1. If an individual transaction belongs to sequence of preceding or subsequent transactions that matches any pattern, than the transaction is labelled as a gaming transaction. Then for each problem-step observed, if any transactions are considered gaming, then the step is labeled as gaming.

incorrect → [guess] & [same answer/diff. context] & incorrect
incorrect → [similar answer] [same context] & incorrect → [similar answer] & [same context] & attempt
incorrect → [similar answer] & incorrect → [same answer/diff. context] & attempt
[guess] & incorrect → [guess] & [diff. answer AND/OR diff. context] & incorrect → [guess] & [diff. answer AND/OR diff. context & attempt
incorrect → [similar answer] & incorrect → [guess] & attempt
help & [searching for bottom-out hint] → incorrect → [similar answer] & incorrect
incorrect → [same answer/diff. context] & incorrect → [switched context before correct] & attempt/help
bug → [same answer/diff. context] & correct → bug
incorrect → [similar answer] & incorrect → [switched context before correct] & incorrect
incorrect → [switched context before correct] & incorrect → [similar answer] & incorrect
incorrect → [similar answer] & incorrect → [did not think before help] & help → incorrect (with first or second answer similar to the last one)
help → incorrect → incorrect → incorrect (with at least one similar answer between steps)
incorrect → incorrect → incorrect → [did not think before help request] & help (at least one similar answer between steps)

Figure B.1: Patterns of Gaming

Table B.1: List of transaction level labels with heuristic thresholds

Identifier	Description
[did not think before help request]	Pause smaller or equal to 5 seconds before a help request
[thought before help request]	Pause greater or equal to 6 seconds before a help request
[read help messages]	Pause greater or equal to 9 seconds per help message after a help request
[scanning help messages]	Pause between 4 and 8 seconds per help message after a help request
[searching for bottom-out hint]	Pause smaller or equal to 3 seconds per help message after a help request
[thought before attempt]	Pause greater or equal to 6 seconds before step attempt
[planned ahead]	Last action was a correct step attempt with a pause greater or equal to 11 seconds
[guess]	Pause smaller or equal to 5 seconds before step attempt
[unsuccessful but sincere attempt]	Pause greater than or equal to 6 seconds before a bug
[guessing with values from problem]	Pause smaller than or equal to 5 seconds before a bug
[read error message]	Pause greater than or equal to 9 seconds after a bug
[did not read error message]	Pause smaller than or equal to 8 seconds after a bug
[thought about error]	Pause greater than or equal to 6 seconds after an incorrect step attempt
[similar answer]	Answer was similar to the previous action (Levenshtein distance of 1 or 2)
[switched context before right]	Context of the current action is not the same as the context for the previous (incorrect) action
[same context]	Context of the current action is the same as the previous action
[repeated step]	Answer and context are the same as the previous action
[diff. answer AND/OR diff. context]	Answer or context is not the same as the previous action

Appendix C

Appendix C: Motivational Surveys

In this section, all of the survey measures that were utilized in the data collected for this study are included below.

C.1 Math Interest [37]

On a scale of 1 to 5, where 1 = "Not at all true" and 5 = "Very true", please indicate the extent to which you agree or disagree with each of the following statements by writing the number that corresponds to your opinion.

1. Math is practical for me to know
2. Math helps me in my daily life outside of school
3. It is important to me to be a person who reasons mathematically
4. Thinking mathematically is an important part of who I am
5. I enjoy the subject of math
6. I like math
7. I enjoy doing math
8. Math is exciting for me

C.2 Self-Efficacy [24]

On a scale of 1 to 9, where 1 = "Not at all" and 9 = "Completely", please indicate the extent to which you agree or disagree with each of the following statements by writing the number that corresponds to your opinion.

1. I am confident that I will do well in math class
2. I expect to do well in math
3. I am confident that I can learn future math concepts
4. Considering the difficulty of this course, I think I will do well in mathematics in the future

5. I am confident that I will do an excellent job on future math problems.

C.3 Achievement Goals [10]

The following statements concern your attitudes toward learning and performance in this class. Please respond to the following items by indicating the degree to which the statement is true of you. Your rating should be on a 7-point scale where 1 = Strongly Disagree and 7 = Strongly Agree.

1. My aim is to completely master the material presented in this unit.
2. In this unit, I am striving to do well compared to other students.
3. In this unit, my goal is to learn as much as possible.
4. In this unit, my aim is to perform well relative to other students.
5. In this unit, my goal is to avoid performing poorly compared to others.
6. I am striving to understand the content of this unit as thoroughly as possible
7. My goal is to perform better than the other students in this unit
8. In this unit, I am striving to avoid performing worse than others.
9. In this unit, my aim is to avoid doing worse than other students.

C.4 Theory of Intelligence [86]

Read the sentences below and then click the one number that shows how much you agree with it. There are no right or wrong answers.

Strongly Agree	Agree	Somewhat Agree	Somewhat Disagree	Disagree	Strongly Disagree
1	2	3	4	5	6

1. You have a certain amount of intelligence and you really can't do too much to change it.
2. Your intelligence is something about you that you can't change very much.
3. You can learn new things, but you can't really change your intelligence.
4. No matter who you are, you can change your intelligence a lot.
5. You can always greatly change how intelligent you are.
6. No matter how much intelligence you have, you can always change it quite a bit.

C.5 Effort Regulation [3]

Please rate the following items based on your behavior in math class. Your rating should be on a 7-point scale where 1 = not at all true of me to 7 = very true of me.

1. I often feel so lazy or bored when I do homework for this class that I quit before I finish what I planned to do.
2. I work hard to do well in this class even if I don't like what we are doing.
3. When class work is difficult, I give up or only study the easy parts.
4. Even when class materials are dull and uninteresting, I manage to keep working until I finish.

Appendix D

Appendix D: Additional Study Result Details

D.1 Classroom and Social Factors

D.1.1 Construct Validity Tables

Table D.1: Pearson's R relating Class Information Behavior measures with motivation measures

	Total time	Percent Opportunity	Avg Session Length	Start Speed
Effort Reg.	0.098	0.186**	0.216***	-0.248***
Math Interest	0.097	0.194**	0.149*	-0.209**
Self-Efficacy	-0.016	0.121**	0.148**	-0.228***
Mastery App.	-0.061	0.181*	0.177**	-0.236***
Perf. App.	-0.029	0.152*	0.207**	-0.254***
Perf. Avoid.	-0.134*	0.026	0.084	-0.111
Growth Mindset	-0.079	0.031	-0.021	-0.023

* - $p < 0.05$, ** - $p<0.01$, *** - $p<0.001$

Table D.2: Pearson's R relating social information behavior measures with motivation measures

	Avg Session Time	Rel Session Time	Start Speed	Relative Start Speed
Effort Reg.	0.216**	0.156*	-0.247***	-0.145*
Math Interest	0.155*	0.096	-0.208*	-0.103
Self-Efficacy	0.182*	0.101	-0.235***	-0.068
Mastery App.	0.213**	0.099	-0.253***	-0.108
Perf. App.	0.089**	-0.022	-0.111***	0.024
Perf. Avoid.	0.153	0.092	-0.226	-0.110
Growth Mindset	-0.031	0.046	-0.027	-0.0002

* - $p < 0.05$, ** - $p<0.01$, *** - $p<0.001$

D.2 Simulation

D.2.1 Instrument Correlation with Diligence

Table D.3: r^2 comparing multi-operational diligence measures with latent diligence parameter for simulated student data

	Baseline	Self-Efficacy	Intrinsic Interest	Combined
Opportunity	0.50	0.40	0.28	0.23
Off-Task	0.34	0.40	0.24	0.24
Start speed	0.41	0.30	0.08	0.17
Early Finish	0.34	0.20	0.25	0.15
Opp, Off-Task	0.57	0.61	0.36	0.37
Opp, Start	0.55	0.45	0.19	0.23
Opp, Finish	0.44	0.31	0.29	0.21
Off-Task, Start	0.52	0.52	0.20	0.29
Off-Task, Finish	0.50	0.50	0.41	0.33
Start, Finish	0.58	0.44	0.25	0.25
Opp, Off-Task, Start	0.61	0.57	0.24	0.29
Opp, Off-Task, Finish	0.53	0.40	0.27	0.26
Opp, Start, Finish	0.53	0.40	0.27	0.25
Off-Task, Start, Finish	0.62	0.60	0.28	0.33
Opp, Off-Task, Start, Finish	0.60	0.49	0.31	0.29

D.3 Multi-operationalization of Diligence with Student Data

Table D.4: Latent Factor Correlations with Achievement Measures

	Prior Final Grade	Final Grade	Units Completed
F0 (Diligence)	0.252	0.320	0.544
F1 (Procrastinator)	-0.113	-0.141	-0.355
F2 (Slacker)	-0.111	-0.143	-0.121
Attendance	0.137	0.306	0.571
Pct Opportunity	0.195	0.360	0.645
Gaming	-0.138	-0.292	-0.342
Start Speed	-0.239	-0.303	-0.551
Relative Start Speed	-0.242	-0.334	-0.480
Early Finish	-0.111	-0.144	-0.123
Relative Early Finish	-0.038	-0.158	-0.024
Pct Session	0.133	0.189	0.537
Relative Pct Session	0.259	0.409	0.496

Table D.5: Latent Factor Correlations with Motivation Measures

	Effort Reg.	Math Interest	Self-Efficacy	Growth Mindset
F0 (Diligence)	0.243	0.224	0.245	0.024
F1 (Procrastinator)	-0.103	-0.115	-0.139	0.004
F2 (Slacker)	-0.113	-0.082	-0.078	-0.025
Attendance	0.100	0.102	0.024	0.047
Pct Opportunity	0.190	0.198	0.124	0.033
Gaming	-0.177	-0.183	-0.136	-0.042
Start Speed	-0.230	-0.218	-0.243	-0.018
Relative Start Speed	-0.133	-0.108	-0.120	0.004
Early Finish	-0.114	-0.083	-0.078	-0.025
Relative Early Finish	-0.050	0.003	-0.020	-0.136
Pct Session	0.212	0.174	0.183	-0.026
Relative Pct Session	0.167	0.117	0.125	0.074

Table D.6: Latent Factor Correlations with Achievement Goal Measures

	Mastery App.	Perf. App.	Perf. Avoid
F0 (Diligence)	0.254	0.261	0.127
F1 (Procrastinator)	-0.077	-0.116	-0.028
F2 (Slacker)	-0.0152	-0.117	-0.087
Attendance	0.074	0.008	-0.042
Pct Opportunity	0.183	0.155	0.028
Gaming	-0.107	-0.068	-0.056
Start Speed	-0.231	-0.249	-0.113
Relative Start Speed	-0.062	-0.103	0.024
Early Finish	-0.153	-0.118	-0.087
Relative Early Finish	-0.077	0.001	-0.022
Pct Session	0.180	0.213	0.089
Relative Pct Session	0.100	0.100	-0.030

Table D.7: Reliability and Stability of Diligence Measures

Measure	Stability	Reliability
F0 (Diligence)	0.71	0.80
F1 (Procrastinator)	0.36	0.38
F2 (Slacker)	0.39	0.61
Attendance	0.43	0.40
Pct Opportunity	0.65	0.79
Start Speed	0.55	0.64
Relative Start Speed	0.64	0.71
Early Finish	0.34	0.46
Relative Early Finish	0.27	0.34
Pct Session	0.67	0.74
Relative Pct Session	0.63	0.78

Appendix E

Appendix E: Linked Resources

E.1 Datasets

The dataset collected from [47] can be accessed at this address:

https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=613

The simulated student data set that was generated for the analysis in 6 can be accessed at this address:

https://pslcdatashop.web.cmu.edu/Project?id=806

E.2 Code and Libraries

Code for the LEnS framework, simulation runs, and data analysis performed in 6 can be accessed at this address:

https://github.com/stevencdang/MotivSim

E.3 Published Work

The work described in chapter 3 was originally presented at the 2019 International Conference for Educational Data Mining and published in the associated proceedings [83]. The work described in chapter 4 was originally presented at the 2020 International Conference for Educational Data Mining and published in the associated proceedings [85].

Bibliography

[1] John L Horn. “A Rationale and Test for the Number of Factors in Factor ANalysis”. In: 30.2 (1965), pp. 179–185. doi: https://doi.org/10.1007/BF02289447.

[2] Walter Mischel, Yuichi Shoda, and Monica L. Rodriguez. “Delay of gratification in children”. In: Science 244.4907 (May 1989), pp. 933–938. doi: 10.1126/science.2658056.

[3] Paul R Pintrich. “A Manual for the Use of the Motivated Strategies for Learning Questionnaire (MSLQ).” In: (1991).

[4] Martin V Covington. Making the grade: A self-worth perspective on motivation and school reform. New York, NY, US: Cambridge University Press, 1992. doi: 10.1017/CBO9781139173582.

[5] Domenic V Cicchetti. “Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology”. In: Psychological Assessment 6.4 (1994), pp. 284–290. doi: 10.1037/1040-3590.6.4.284.

[6] Albert T. Corbett and John R. Anderson. “Knowledge tracing: Modeling the acquisition of procedural knowledge”. In: User Modelling and User-Adapted Interaction 4.4 (Dec. 1994), pp. 253–278. doi: 10.1007/BF01099821.

[7] Allan Wigfield et al. “Change in children’s competence beliefs and subjective task values across the elementary school years: A 3-year study.” In: Journal of Educational Psychology 89.3 (1997), pp. 451–469.

[8] Dale H Schunk and Peggy A Ertmer. “Self-regulatory processes during computer skill acquisition: Goal and self-evaluative influences.” In: Journal of Educational Psychology 91.2 (1999), pp. 251–260. doi: 10.1037/0022-0663.91.2.251.

[9] Arthur A Stone et al. The Science of Self-report: Implications for Research and Practice. New York: Psychology Press, 1999. doi: 10.4324/9781410601261.

[10] Paul R Pintrich. “An Achievement Goal Theory Perspective on Issues in Motivation Terminology, Theory, and Research”. In: Contemporary Educational Psychology 25.1 (2000), pp. 92–104. doi: 10.1006/ceps.1999.1017.

[11] Allan Wigfield and Jacquelynne S Eccles. “Expectancy–Value Theory of Achievement Motivation”. In: Contemporary Educational Psychology 25.1 (Jan. 2000), pp. 68–81.

[12] Judith M Harackiewicz et al. “Revision of achievement goal theory: Necessary and illuminating.” In: Journal of Educational Psychology 94.3 (2002), pp. 638–645. doi: https://doi.org/10.1037/0022-0663.94.3.638.

[13] Steven J Heine et al. “What’s wrong with cross-cultural comparisons of subjective Likert scales?: The reference-group effect.” In: Journal of Personality and Social Psychology 82.6 (2002), pp. 903–918. doi: 10.1037/0022-3514.82.6.903.

[14] Sander Nieuwenhuis and Stephen Monsel. “Residual costs in task switching: Testing the failure-to-engage hypothesis”. In: Psychonomic Bulletin and Review 9.1 (Mar. 2002), pp. 86–92. doi: 10.3758/BF03196259.

[15] Steven J Scher and Nicole M Osterman. “Procrastination, conscientiousness, anxiety, and goals: Exploring the measurement and correlates of procrastination among school-aged children”. In: Psychology in the Schools 39.4 (2002), pp. 385–398. doi: https://doi.org/10.1002/pits.10045.

[16] Robert J Marzano, Jana S Marzano, and Debra Pickering. Classroom management that works: Research-based strategies for every teacher. Research-based strategies for every teacher. 2003.

[17] Christopher A Wolters. “Understanding procrastination from a self-regulated learning perspective.” In: Journal of Educational Psychology 1.95 (2003), p. 179.

[18] Ryan Shaun Baker, Albert T Corbett, and Kenneth R Koedinger. “Detecting Student Misuse of Intelligent Tutoring Systems”. In: International Conference on Intelligent Tutoring Systems. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 531–540. doi: 10.1007/978-3-540-30139-4_50.

[19] Ryan Shaun Baker et al. “Off-Task Behavior in the Cognitive Tutor Classroom: When Students “Game the System””. In: Proceedings of the 2004 Conference on Human Factors in Computing Systems - CHI ’04. New York, New York, USA: ACM Press, Apr. 2004, pp. 383–390. doi: 10.1145/985692.985741.

[20] Christopher A Wolters. “Advancing Achievement Goal Theory: Using Goal Structures and Goal Orientations to Predict Students’ Motivation, Cognition, and Achievement.” In: Journal of Educational Psychology 96.2 (2004), pp. 236–250. doi: 10.1037/0022-0663.96.2.236.

[21] Jacquelynne S Eccles. “Subjective Task Value and the Eccles et al. Model of Achievement-Related Choices.” In: Handbook of Competence and Motivation. Ed. by Andrew J Elliot and Carol S Dweck. New York, NY, US: Guilford Publications, 2005, pp. 105–121.

[22] Vincent Aleven et al. “Toward Meta-cognitive Tutoring: A Model of Help Seeking with a Cognitive Tutor”. In: International Journal of Artificial Intelligence in Education 16.2 (2006), pp. 101–128.

[23] Ryan S. Baker, Albert T Corbett, and Angela Z Wagner. “Human Classification of Low-Fidelity Replays of Student Actions”. In: 2002 (May 2006), pp. 29–36.

[24] Albert Bandura et al. “Guide for constructing self-efficacy scales”. In: Self-efficacy beliefs of adolescents 5.1 (2006), pp. 307–337.

[25] Ryan S Baker. “Is gaming the system state-or-trait? Educational data mining through the multi-contextual application of a validated behavioral model”. In: Complete On-Line Proceedings of the Workshop on Data Mining for User Modeling at the 11th International Conference on User Modeling 2007. Vol. 2007. 2007, pp. 76–80.

[26] Brandon J Schmeichel. Attention control, memory updating, and emotion regulation temporarily reduce the capacity for executive control. 2007. doi: 10.1037/0096-3445.136.2.241.

[27] Piers Steel. "The nature of procrastination: a meta-analytic and theoretical review of quintessential self-regulatory failure." In: Psychological Bulletin 133.1 (Jan. 2007), pp. 65–94. doi: 10.1037/0033-2909.133.1.65.

[28] Ryan S. Baker and Adriana de Carvalho. "Labeling student behavior faster and more precisely with text replays". In: Educational Data Mining 2008. 2008.

[29] Ryan S. Baker et al. "Why students engage in "gaming the system" behavior in interactive learning environments". In: Journal of Interactive Learning Research 19.2 (2008), pp. 185–224.

[30] Chris S Hulleman et al. "Task values, achievement goals, and interest: An integrative analysis." In: Journal of educational psychology 100.2 (Sept. 2008), pp. 398–416. doi: 10.1037/0022-0663.100.2.398.

[31] Oboru Matsuda et al. "A Computational Model of How Learner Errors Arise from Weak Prior Knowledge". In: Proc. of the Annual Conference of the Cognitive Science Society. 2009, pp. 1288–1293.

[32] Jonathan P Rowe et al. "Off-Task Behavior in Narrative-Centered Learning Environments". In: Proceeedings of the 14th International Conference on AI in Education. 2009, pp. 99–106.

[33] Rafael A. Calvo and Sidney D'Mello. "Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications". In: IEEE Transactions on Affective Computing 1.1 (Jan. 2010), pp. 18–37. doi: 10.1109/T-AFFC.2010.1.

[34] Avi Kaplan and Hanoch Flum. "Achievement goal orientations and identity formation styles". In: Educational Research Review 5.1 (2010), pp. 50–67. doi: https://doi.org/10.1016/j.edurev.2009.06.004.

[35] Kenneth R Koedinger et al. "A data repository for the EDM community: The PSLC DataShop". In: Handbook of educational data mining 43 (2010), pp. 43–56.

[36] Wouter Kool et al. "Decision Making and the Avoidance of Cognitive Demand". In: Journal of Experimental Psychology: General 139.4 (2010), pp. 665–682. doi: 10.1037/a0020198.

[37] Lisa Linnenbrink-Garcia et al. "Measuring Situational Interest in Academic Domains". In: Educational and Psychological Measurement 70.4 (Apr. 2010), pp. 647–671. doi: 10.1177/0013164409355699.

[38] Bernard Weiner. "The Development of an Attribution-Based Theory of Motivation: A History of Ideas". In: Educational Psychologist 45.1 (Jan. 2010), pp. 28–36. doi: 10.1080/00461520903433596.

[39] Kasia Muldner et al. "An analysis of students' gaming behaviors in an intelligent tutoring system: predictors and impacts". In: User Modeling and User-Adapted Interaction. Vol. 21. 1-2. Jan. 2011, pp. 99–135. doi: 10.1007/s11257-010-9086-0.

[40] Jennifer Sabourin, Bradford Mott, and James C Lester. "Generalizing Models of Student Affect in Game-Based Learning Environments". In: Affective Computing and

Intelligent Interaction. Berlin, Heidelberg: Springer, Berlin, Heidelberg, Oct. 2011, pp. 588–597. doi: 10.1007/978-3-642-24571-8_73.

[41] Dale H Schunk and Barry J Zimmerman, eds. Handbook of self-regulation of learning and performance. Educational psychology handbook series. New York, NY, US: Routledge, 2011.

[42] Corwin Senko, Chris S Hulleman, and Judith M Harackiewicz. "Achievement Goal Theory at the Crossroads: Old Controversies, Current Challenges, and New Directions". In: Educational Psychologist 46.1 (Jan. 2011), pp. 26–47.

[43] Ryan SJd Baker et al. "Towards Sensor-Free Affect Detection in Cognitive Tutor Algebra." In: Proceedings of the 5th International Conference on Educational Data Mining. (2012).

[44] Michel C. Desmarais and Ryan S J D Baker. "A review of recent advances in learner and skill modeling in intelligent learning environments". In: User Modelling and User-Adapted Interaction 22 (2012), pp. 9–38. doi: 10.1007/s11257-011-9106-8.

[45] Kenneth R Koedinger, Albert T Corbett, and Charles Perfetti. "The Knowledge-Learning-Instruction Framework: Bridging the Science-Practice Chasm to Enhance Robust Student Learning". In: Cognitive Science 36.5 (2012), pp. 757–798. doi: 10.1111/j.1551-6709.2012.01245.x. url: http://search.library.cmu.edu/vufind/Summon/Record?id=FETCH-eric_primary_EJ9721101.

[46] Carlos J Asarta and James R Schmidt. "Access patterns of online materials in a blended course". In: Decision Sciences Journal of Innovative Education 11.1 (2013), pp. 107–123. doi: https://doi.org/10.1111/j.1540-4609.2012.00366.x.

[47] Matthew L Bernacki and Steven Ritter. Hopewell 2011-2012. Dataset 613 in Datashop. https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=613. 2013.

[48] Angela Lee Duckworth and Stephanie M. Carlson. "Self-Regulation and School Success". In: Self-Regulation and Autonomy: Social and Developmental Dimensions of Human Conduct. Ed. by Bryan W. Sokol, Frederick M. E. Grouzet, and UlrichEditors Müller. Cambridge University Press, 2013, pp. 208–230. doi: 10.1017/CBO9781139152198.015.

[49] Janice D Gobert et al. "From Log Files to Assessment Metrics: Measuring Students' Science Inquiry Skills Using Educational Data Mining". In: Journal of the Learning Sciences 22.4 (Oct. 2013), pp. 521–563. doi: 10.1080/10508406.2013.837391.

[50] John S Kinnebrew, Kirk M Loretz, and Gautam Biswas. "A contextualized, differential sequence mining method to derive students' learning behavior patterns". In: JEDM| Journal of Educational Data Mining 5.1 (2013), pp. 190–219. doi: 10.5281/zenodo.3554617.

[51] Christine Otieno et al. "Can Help Seeking Behavior in Intelligent Tutoring Systems Be Used as Online Measure for Goal Orientation?" In: Annual Meeting of the Cognitive Science Society. Vol. 35. 35. 2013.

[52] Nicole Shechtman et al. Promoting grit, tenacity, and perseverance: Critical factors for success in the 21st century. Washington D.C., 2013.

[53] Matthew L Bernacki, Vincent Aleven, and Timothy J Nokes-Malach. "Stability and change in adolescents' task-specific achievement goals and implications for learning

mathematics with intelligent tutors". In: Computers in Human Behavior 37 (Aug. 2014), pp. 73–80. doi: 10.1016/j.chb.2014.04.009.

[54] Stephen E Fancsali et al. "Goal Orientation, Self-Efficacy, and "Online Measures" in Intelligent Tutoring Systems." In: Proceedings of the Annual Meeting of the Cognitive Science Society 36.36 (2014).

[55] Brian M Galla et al. "The Academic Diligence Task (ADT): assessing individual differences in effort on tedious but important schoolwork". In: 39.4 (Oct. 2014), pp. 314–325.

[56] Jay H Hardy III. "Dynamics in the self-efficacy–performance relationship following failure". In: Personality and Individual Differences 71 (Dec. 2014), pp. 151–158. doi: 10.1016/j.paid.2014.07.034.

[57] Caitlin Mills, Nigel Bosch, and Art Graesser. "To Quit or Not to Quit: Predicting Future Behavioral Disengagement from Reading Patterns". In: International Conference on Intelligent Tutoring Systems. Cham: Springer, June 2014, pp. 19–28. doi: 10.1007/978-3-319-07221-0_3.

[58] Luc Paquette, Adriana M J A de Carvalho, and Ryan S Baker. "Towards Understanding Expert Coding of Student Disengagement in Online Learning". In: Proceedings of the 36th Annual Meeting of the Cognitive Science Society. 2014, pp. 1126–1131.

[59] J Elizabeth Richey et al. "Relating a Task-Based, Behavioral Measure of Achievement Goals to Self-Reported Goals and Performance in the Classroom". In: Proceedings of the Annual Meeting of the Cognitive Science Society. Vol. 36. 36. 2014.

[60] Shuhua Sun, Jeffrey B Vancouver, and Justin M Weinhardt. "Goal choices and planning: Distinct expectancy and value effects in two goal processes". In: Organizational Behavior and Human Decision Processes 125.2 (Nov. 2014), pp. 220–233. doi: 10.1016/j.obhdp.2014.09.002.

[61] Bernacki, Matthew L, Nokes-Malach, Timothy J, and Aleven, Vincent. "Examining self-efficacy during learning: variability and relations to behavior, performance, and learning". In: Metacognition and Learning 10.1 (2015), pp. 99–117. doi: 10.1007/s11409-014-9127-x.

[62] Matthew Botvinick and Todd Braver. "Motivation and Cognitive Control: From Behavior to Neural Mechanism". In: Annual Review of Psychology 66.1 (Jan. 2015), pp. 83–113. doi: 10.1146/annurev-psych-010814-015044.

[63] Angela L Duckworth and David Scott Yeager. "Measurement Matters: Assessing Personal Qualities Other Than Cognitive Ability for Educational Purposes". In: Educational Research 44.4 (May 2015), pp. 237–251. doi: 10.3102/0013189X15584327.

[64] Jessica Kay Flake et al. "Measuring cost: The forgotten component of expectancy-value theory". In: Contemporary Educational Psychology 41 (Apr. 2015), pp. 232–244. doi: 10.1016/j.cedpsych.2015.03.002.

[65] Christian Köppe et al. "Flipped classroom patterns: designing valuable in-class meetings". In: Proceedings of the 20th European Conference on Pattern Languages of Programs. 2015, pp. 1–17. doi: 10.1145/2855321.2855348.

[66] Caitlin Mills, Nigel Bosch, and Andrew M Olney. "Mind Wandering During Learning with an Intelligent Tutoring System". In: International conference on artificial

intelligence in education. Vol. 9112. Cham: Springer, June 2015, pp. 267–276. doi: 10.1007/978-3-319-19773-9_27.

[67] Jaclyn Ocumpaugh, Ryan S Baker, and Ma Mercedes T Rodrigo. Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 Technical and Training Manual. Tech. rep. Feb. 2015.

[68] Amy Ogan et al. "Towards Understanding How to Assess Help-Seeking Behavior Across Cultures". In: International Journal of Artificial Intelligence in Education 25.2 (2015), pp. 229–248. doi: 10.1007/s40593-014-0034-8.

[69] Luc Paquette et al. "Cross-System Transfer of Machine Learned and Knowledge Engineered Models of Gaming the System". In: User Modeling, Adaptation and Personalization - 23rd International Conference, UMAP 2015, Proceedings. Ed. by Kalina Bontcheva et al. Springer, Apr. 2015, pp. 183–194. doi: 10.1007/978-3-319-20267-9_15.

[70] Valerie J Shute et al. "Modeling how incoming knowledge, persistence, affective states, and in-game progress influence student learning from an educational game". In: Computers & Education 86 (Aug. 2015), pp. 224–235. doi: 10.1016/j.compedu.2015.08.001.

[71] Shayan Doroudi et al. "Sequence Matters, But How Exactly? A Method for Evaluating Activity Sequences from Data". In: Proceedings of the 9th International Conference on Educational Data Mining. 2016, pp. 70–77.

[72] Angela L Duckworth, Tamar Szabó Gendler, and James J Gross. "Situational Strategies for Self-Control". In: Perspectives on Psychological Science 11.1 (Jan. 2016), pp. 35–55. doi: 10.1177/1745691615623247.

[73] Stephen Hutt et al. "The Eyes Have It: Gaze-Based Detection of Mind Wandering during Learning with an Intelligent Tutoring System." In: International Educational Data Mining Society (2016).

[74] Christopher J Maclellan et al. "The Apprentice Learner architecture: Closing the loop between learning theory and educational data". In: Proceedings of the 9th International Conference on Educational Data Mining. 2016, pp. 151–158.

[75] Angela Stewart et al. "Where's Your Mind At?: Video-Based Mind Wandering Detection During Film Viewing". In: 2016 Conference on User Modeling, Adaptation, and Personalization. ACM, July 2016, pp. 295–296. doi: 10.1145/2930238.2930266.

[76] Elliot T Berkman et al. "Self-Control as Value-Based Choice". In: Current Directions in Psychological Science 26.5 (Oct. 2017), pp. 422–428. doi: 10.1177/0963721417704394.

[77] Sidney D'Mello, Ed Dieterie, and Angela Duckworth. "Advanced, Analytic, Automated (AAA) Measurement of Engagement During Learning". In: Educational Psychologist 52.2 (Feb. 2017), pp. 104–123. doi: 10.1080/00461520.2017.1281747.

[78] Luc Paquette and Ryan S Baker. "Variations of Gaming Behaviors Across Populations of Students and Across Learning Environments". In: Artificial Intelligence in Education. Ed. by Elisabeth André et al. Cham: Springer, 2017, pp. 274–286.

[79] Angela Stewart et al. "Face Forward: Detecting Mind Wandering from Video During Narrative Film Comprehension". In: Artificial Intelligence in Education. Cham: Springer, June 2017, pp. 359–370.

[80] Cristina E Dumdumaya, Michelle P Banawan, and Ma Mercedes T Rodrigo. “Identifying Students’ Persistence Profiles in Problem Solving Task”. In: Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization. UMAP ’18. New York, New York, USA: ACM Press, 2018, pp. 281–286. doi: 10.1145/3213586.3225237.

[81] Shamya Karumbaiah, Ryan S Baker, and Valerie Shute. “Predicting Quitting in Students Playing a Learning Game.” In: International Educational Data Mining Society (2018).

[82] Shamya Karumbaiah, Ryan S. Baker, and Valerie Shute. “Predicting Quitting in Students Playing a Learning Game”. In: Proceedings of the 11th International Conference on Educational Data Mining. Mar. 2018. url: https://pluto.coe.fsu.edu/ppteam/pp-links/.

[83] Steven Dang and Ken Koedinger. “Exploring the Link Between Motivations and Gaming”. In: Proceedings of the 12th International Conference on Educational Data Mining. 2019, pp. 276–281.

[84] Huggins-Manley A Corinne et al. “A Commentary on Construct Validity When Using Operational Virtual Learning Environment Data in Effectiveness Studies”. In: Journal of Research on Educational Effectiveness 12.4 (Jan. 2020), pp. 750–759. doi: 10.1080/19345747.2019.1639869.

[85] Steven C Dang and Kenneth R Koedinger. “The Ebb and Flow of Student Engagement Measuring motivation through temporal pattern of self-regulation”. In: Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020). 2020, pp. 61–68.

[86] Lisa S Blackwell, Kali H Trzesniewski, and Carol Sorich Dweck. “Implicit theories of intelligence predict achievement across an adolescent transition: a longitudinal study and an intervention.” In: Child development 78.1 (), pp. 246–63. doi: 10.1111/j.1467-8624.2007.00995.x.