Exploring Dynamical Assessments of Affect, Behavior, and Cognition and Math State Test Achievement

Maria Ofelia Z. San Pedro[1], Erica L. Snow[2], Ryan S. Baker[1], Danielle S. McNamara[2], Neil T. Heffernan[3]
[1]Teachers College Columbia University, 525 W 120th St. New York, NY 10027
[2]Arizona State University, Learning Sciences Institute, 1000 S. Forest Mall, Tempe, AZ 85287
[3]Worcester Polytechnic Institute, 100 Institute Rd. Worcester, MA 01609

mzs2106@tc.columbia.edu, Erica.L.Snow@asu.edu, baker2@exchange.tc.columbia.edu, Danielle.McNamara@asu.edu, nth@wpi.edu

ABSTRACT
There is increasing evidence that fine-grained aspects of student performance and interaction within educational software are predictive of long-term learning. Machine learning models have been used to provide assessments of affect, behavior, and cognition based on analyses of system log data, estimating the probability of a student's particular affective state, behavior, and knowledge (cognition). These measures have (in aggregate) successfully predicted outcomes such as performance on standardized exams. In this paper, we employ a different approach of relating interaction patterns to learning outcomes, using dynamical methods that assess patterns of fine-grained measures of affect, behavior, and knowledge as they occur across time. We use Hurst exponents and Entropy scores computed from assessments of affect, behavior, performance, and knowledge acquired from 1,376 middle school students who used a math tutoring system (ASSISTments), and analyze the relations of these dynamical measures to the students' end-of-year state test (MCAS) performance. Our results show that fine-grained changes in affect, behavior, and knowledge are significantly related to and predictive of their eventual MCAS performance, providing a new lens on the dynamic and nuanced nature of student interaction within online learning platforms and how it affects achievement.

Keywords
Affect Detection, Knowledge Modeling, Educational Data Mining, Hurst, Entropy

1. INTRODUCTION
The increasing deployment of educational software in classrooms has provided new opportunities for studying a broad range of student modeling constructs. The ability of these systems to log student interaction in fine-grained detail has led to the development of automated detectors or models of student learning and engagement [1, 4, 5, 6, 10]. It has been demonstrated through *discovery with models* analyses [20] that detector assessments of engagement and learning can be used to predict long-term student outcomes such as performance in end-of-year standardized exams [24], college enrollment [31] and college major choice [33], even several years after the student engages in online learning. The fine-grained measures of learning and engagement at the action level are then aggregated at the student-level in forming a training dataset for the prediction of learning outcomes. However, these assessments often use simple aggregation methods such as student-level averages, whereas it is known that there are complex patterns in how affect develops over time (e.g. [14]). Hence these simple methods of aggregation may miss fine-grained and nuanced patterns in affect or behaviors that manifest across time.

Indeed, research has also shown that students' learning behaviors are complex and dynamic in nature [19]. Recent work has begun to evaluate *interaction patterns* within learning tasks. This work has revealed that fine-grained pattern analysis can shed light upon various cognitive, behavioral, and learning outcomes [21, 22, 29, 30, 37, 38]. For example, Lee and colleagues [21], and Liu and colleagues [22] evaluated how 3-step sequences of confusion [21, 22] and frustration [22] correlate to learning outcomes. Rodrigo and colleagues [29] also found that 3-step sequences of affective states (boredom, engaged concentration, confusion, and delight) from fine-grained detectors correlated to differences in learning outcomes. Sabourin and colleagues [30] found that the impact of student behavior on learning outcomes depended in part on the affect that preceded the behavior. Results from these studies reveal that fluctuations in students' affect and behavior over time (assessed through automated detectors) play important roles in learning outcomes.

However, much of this work had the limitation of only considering changes over brief periods of time. In this paper, we address this limitation by employing dynamical methodologies to quantify nuanced patterns of student affect, behavior, and learning across time, specifically two academic school years. We utilize fine-grained measures of affect, behavior, and knowledge (cognition) from middle school students who used the ASSISTments systems, and compute dynamical measures (i.e., Hurst and Entropy) of these constructs for each student. These measures (see below for details) characterize the occurrence and type of behavior across time for the constructs of interest (affect, behavior, knowledge) for each student within the ASSISTments environment.

We use two types of dynamical analysis techniques, Entropy and Hurst exponents. Entropy is a statistical measure used to assess the amount of predictability present within a time series [34]. Previously, Entropy has been used in EDM analyses by Snow and colleagues [38], to quantify the amount of randomness in students' interaction patterns within a game-based interface. Using this methodology they found that students who acted in more controlled (and predictable) manners had significantly higher task performance compared to students who acted in more random (or unpredictable) fashions. Hurst exponents are similar to Entropy in that they categorize the amount of order present within a system; however, unlike Entropy, Hurst exponents act as long-term correlations that capture how each moment in a time series

relates to the others. Thus, Hurst provides an even finer-grained look at the emergence of patterns across long periods of time. Recently, Hurst exponents have been used to characterize students' learning behaviors within game-based environments. For instance, Snow and colleagues [36] used this technique to examine nuanced fluctuations in students' choice patterns across time. Using the Hurst exponent, Snow and colleagues again found that students who acted in more deterministic manners (i.e., controlled and planned) were more likely to demonstrate higher learning gains compared to students who acted in more random (or impetuous) manners.

In the current work, we evaluate the degree to which Entropy and Hurst exponent measures based on affect, behavior, and knowledge (cognition) predicts a longer-term outcome, students' end-of-year state exam performance. This research was conducted on a dataset of 1,376 students who used ASSISTments when they were in middle school during the school years of 2004-2005 to 2005-2006 and took the standardized end-of-year state exams. We investigate in particular, the following research questions:

1) How are fluctuations in patterns of students' affect, behavior, and knowledge related to their end-of-year state math achievement test scores?

2) Are dynamical measures of affect, behavior, and knowledge predictive of student performance outcomes (end-of-year test score, i.e., MCAS)?

2. METHODOLOGY

2.1 Data Source: The ASSISTments System

This study explores students' learning outcomes and their interaction patterns from their usage of the ASSISTments system [27], a web-based tutoring system for middle-school mathematics, provided to students for free by Worcester Polytechnic Institute (WPI). As of 2013, ASSISTments has been used by over 50,000 students a year as part of their regular mathematics classes. ASSISTments *assesses* a student's knowledge while *assisting* them in learning, providing teachers with formative assessment of students as they progress in their acquisition of specific knowledge components.

Within the system, each problem maps to one or more cognitive skills. When students who are working on an ASSISTments problem answer correctly, they proceed to the next problem. When they answer incorrectly (Figure 1), the system scaffolds instruction by dividing the problem into component parts, stepping students through each before returning them to the original problem (as in Figure 2). Once the correct answer to the original question is provided, the student is prompted to go to the next question. Teachers use ASSISTments in designing problem sets completed by students either during class time or as homework assignments. ASSISTments provides data on student performance that is used by teachers to track misconceptions and discuss them in class.

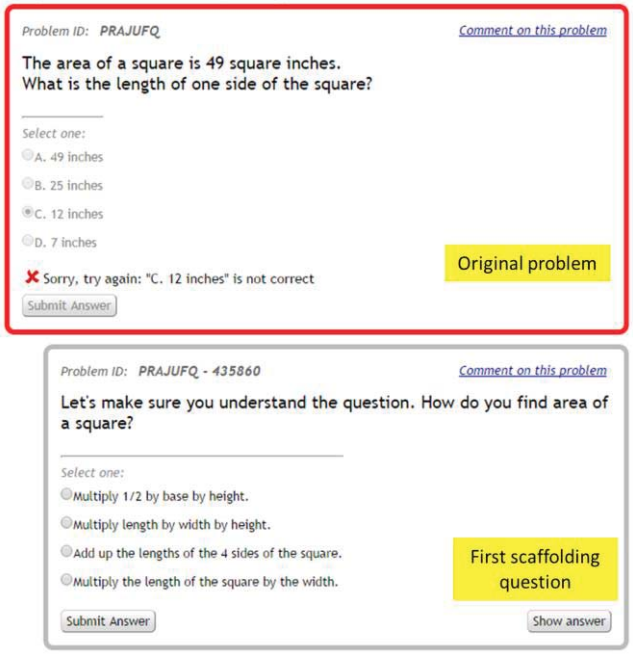

Figure 1. Example of an ASSISTments problem.

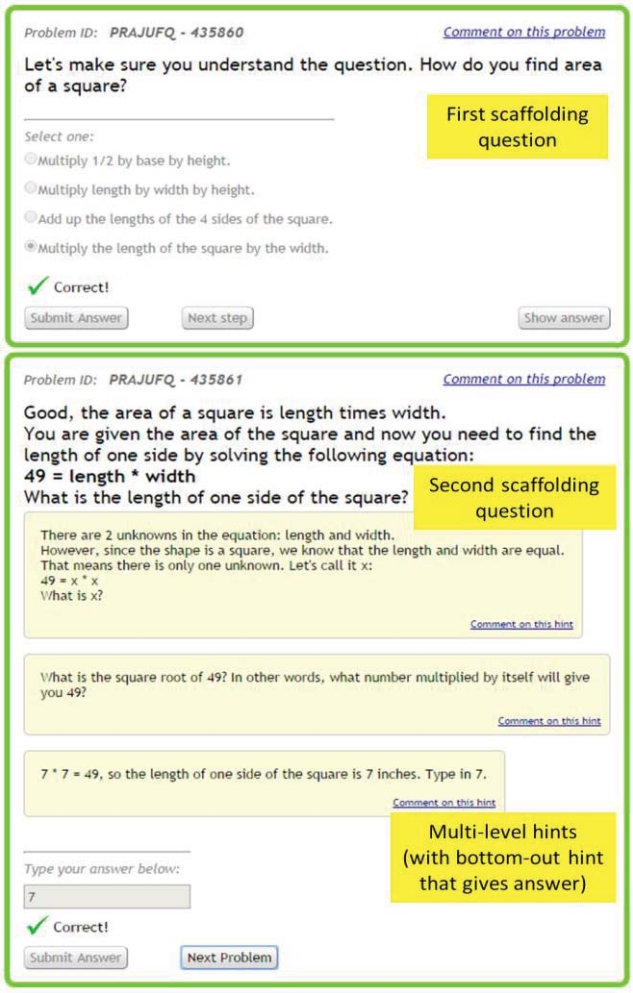

Figure 2. Example of Scaffolding and Hints in an ASSISTments Problem.

2.2 Data

2.2.1 State Exam Scores

Students who used ASSISTments when they were in middle school also took the MCAS (Massachusetts Comprehensive Assessment System) state standardized test near the end of their school years. The test is composed of English Language Arts, Mathematics and Science, and Technology subjects. This study analyzes usage of a tutoring system in mathematics; consequently, we examined the relationship of performance to the MCAS test scores for the math portion. Raw scores for the math portion range from 0 to 54 and are later scaled by the state after all tests have been scored. The scaled scores can be categorized into four groups: Failing, Needs Improvement, Proficient, and Advanced. Students in Massachusetts are required to score above failing to be able to graduate from high school; if students score in the Advanced group, they automatically earn a scholarship to a state college.

2.2.2 ASSISTments Data

Interaction log files from ASSISTments were obtained for 1,376 students who used the system when they were in middle school ranging from school years 2004-2005 to 2005-2006 (these school years were used due to the availability of the state exam data for these particular cohorts). These students, diverse in terms of both ethnicity and socio-economic status, were drawn from middle schools in an urban district in New England who used the ASSISTments system systematically during the school years. The 1,376 students generated a total of 830,167 actions within the system (an action may be answering a question, or requesting help), across around 3,700 original and scaffolding problems from ASSISTments, with an average of approximately 220 ASSISTments problems per student. Affect, behavior, and knowledge models were applied to this dataset to evaluate interaction patterns.

2.3 Computing Interaction Features

The interaction features used to compute dynamical assessments were generated using automated detectors of student engagement and learning previously developed and validated for ASSISTments. These included existing models of educationally-relevant affective states (boredom, engaged concentration, confusion, frustration), disengaged behaviors (off-task behavior and gaming the system), and student knowledge. Each of the detectors was applied to every action in the existing data set, in the same fashion as in previous publications [24]. We also included in our feature set of interactions, information on student correctness over time within ASSISTments.

2.3.1 Affect and Disengaged Behaviors

To obtain assessments of affect and disengaged behaviors, we leveraged existing detectors of student affect and behavior within the ASSISTments system [24]. Detectors of four affective states were utilized: boredom, engaged concentration, confusion, and frustration. Detectors of two disengaged behaviors are utilized: off-task behavior and gaming the system. Because our sample of students came from urban middle schools, their respective data were labeled using models optimized for students in urban schools [23, 24].

The affect and behavior detectors were developed in a two-stage process: first, student affect labels were acquired from field observations conducted using the BROMP protocol and HART Android app (reported in [24]), and then those labels were synchronized with the log files generated by ASSISTments at the same time. This process resulted in automated detectors that can be applied to log files at scale, specifically the data set used in this project (interaction log files for the 1,376 students). The detectors were constructed using only log data from student actions within the software occurring at the same time as or before the observations. The models performed as well as or better than other published models of sensor-free affect detection in educational software [3, 11, 13, 30]. They were then applied to the data set used in this paper to produce confidence values for each construct over time, which were then used to create dynamical assessments of affect and behavior.

2.3.2 Student Knowledge

Corbett and Anderson's [12] Bayesian Knowledge Tracing (BKT) model, a knowledge-estimation model that has been used in a considerable number of online learning systems, was applied to the data for this study. Models were fit by employing brute-force grid search (see [2]). BKT infers students' latent knowledge from their performance on problems that exercise the same set of skills. Each time a student attempts a problem or problem step for the first time, BKT recalculates the estimates of that student's knowledge for the skill (or knowledge component) involved in that problem. Estimations for each skill are made along four parameters: (1) L_O, the initial probability that the student knows the skill, (2) T, the probability of learning the skill at each opportunity to use that skill, (3) G, the probability that the student will give the correct answer despite not knowing the skill, and (4) S, the probability that the student will give an incorrect answer despite knowing the skill. The estimates obtained via BKT were calculated based on the student's first response to each problem, and were applied to each of the student's subsequent attempts on that problem.

We were able to distill interaction features –affect, behavior and knowledge using these models, as well as correctness – for each student action within the ASSISTments system. Affect and behavior features were initially computed at a 20-second grain-size and then applied to all relevant actions. These action-level features values are then used to compute student-level dynamical measures of Hurst and Entropy scores.

2.4 Dynamical Assessments of Student Interaction Features

Variations in students' interaction features (affect, behavior, knowledge, correctness) were assessed using two dynamical methodologies: Entropy analyses and Hurst exponents. These dynamic techniques are used to quantify (in standardized values) variations in students' interaction features and examine how these variations impacted students' year-end standardized test scores (i.e., MCAS). A description and explanation of Entropy analyses and Hurst exponents are described below.

2.4.1 Entropy

Entropy analyses were conducted to quantify the degree to which fluctuations in students' affective states were ordered (i.e., predictable) or disordered (i.e., unpredictable). Entropy analysis is a statistical measure that quantifies the overall tendency (i.e., amount of predictability) of a time series [34]. Entropy has been used across a variety of domains to measure random and ordered processes [15, 17, 34, 35, 38]. In the current study, Entropy is used to gain a deeper understanding of how changes in students' affective states across time may reflect ordered and disordered processes. To calculate Entropy, we applied the affect, behavior, and knowledge series produced from the models discussed above,

to data from school years 2004-2005 and 2005-2006. Entropy was then calculated using the following (standard) formula:

$$H(x) = -\sum_{i=0}^{N} P(x_i)(\log_e P(x_i)) \qquad (1)$$

Within the Entropy equation, $P(x_i)$ represents the probability of a given affective state. For instance, the Entropy for student X is the additive inverse of the sum of products calculated by multiplying the probability of each affect state by the natural log of the probability of that state. This formula affords the ability to capture the degree to which fluctuations in students' affect, behavior, knowledge, and correctness are ordered or disordered.

2.4.2 Hurst

While Entropy provides an overall quantification of a time series, it does not calculate how each moment in the time series may be related to the next. Thus, a more fine-grained analysis is needed to examine how fluctuations in students' affect, behavior, knowledge, and correctness manifest and change across time. To classify the tendency of students' affective states, Hurst exponents were calculated using Detrended Fluctuation Analysis (DFA) [26]. To calculate the Hurst exponent, the DFA integrates the normalized time series and then divides the series into equal intervals of length, *n*. Each interval is then fit with a least squares line and the integrated time series is *detrended* by subtracting the local predicted values (i.e., least square lines for each interval) from the integrated time series. The procedure is repeated for intervals of different lengths, increasing exponentially by the power of 2. Finally, each interval size is assigned a characteristic fluctuation, $F(n)$, that is calculated as the root mean square deviation of the integrated time series from local least squares lines. $\text{Log}_2 F(n)$ is then regressed onto $\log_2(n)$; which produces the slope of the regression line or Hurst exponent, *H*. Hurst exponents range from 0 to 1 and can be interpreted as follows: $0.5 < H \leq 1$ indicates persistent (controlled) behavior, $H = 0.5$ signifies random (independent) behavior, and $0 \leq H < 0.5$ denotes anti-persistent (reversion to the mean) behavior.

2.5 Predictive Modeling of State Test Scores

Prior work has shown that student usage choices while receiving tutoring in ASSISTments can predict as much of the variance in students' end-of-year state test scores as student performance can on items designed to assess test-related knowledge [16, 28]. It has also been shown that machine-learned and fine-grained assessments of affect and behavior can improve predictions of test score performance [24]. We extend this further and explore the value of also understanding the role of the degree of order/disorder of interaction (through occurrences of affect, behavior, knowledge, and correctness) in predicting student learning outcomes as reflected by students' end-of-year standardized examination scores.

After obtaining the aggregate student-level Hurst and Entropy scores for each student's patterns of affect, behavior, knowledge, and correctness, we examined how the degree of variation in the students' interaction patterns within ASSISTments was related to their MCAS math performance. We further examined these relations by conducting linear regression analyses on the students' MCAS math performance. We fit a cross-validated (6-fold, student-level) machine-learned model using linear regression with M5' feature selection to examine how students' dynamical assessments of interaction were predictive of their MCAS math scores. We generated reduced linear regression models that used three feature sets: (1) Hurst scores of interaction only, (2) Entropy scores of interaction only, and (3) both Hurst and Entropy scores of interaction. We then compared their cross-validated model performances and evaluated the features in the model with best performance values.

3. RESULTS

3.1 Hurst, Entropy, and State Test Scores

We first explore the relations between the MCAS scores for math and students' interaction patterns (i.e., their Hurst and Entropy scores) by examining the graphs of student proficiency (from MCAS performance) and the corresponding trends in Hurst and Entropy values. We grouped the students according to their scaled score groupings of Failing, Needs Improvement, Proficient, and Advanced, then computed for the average values of their Hurst and Entropy scores for affect, behavior, knowledge, and correctness in ASSISTments.

The graph of test proficiency and entropy measures (Figure 3) shows that low-achieving and high-achieving students experience fluctuations in affect, behavior, knowledge, and correctness while using ASSISTments in varying degrees. Students who have higher MCAS scores (i.e., *Advanced*) exhibited less fluctuation (lower entropy score) in their frustration ($F(3,1372) = 56.009$, $p < 0.001$, adjusted $\alpha = 0.013$), engaged concentration ($F(3,1372) = 27.334$, $p < 0.001$, adjusted $\alpha = 0.023$), off-task behavior ($\chi^2(3) = 64.089$, $p < 0.001$, adjusted $\alpha = 0.030$), and gaming the system ($\chi^2(3) = 238.350$, $p < 0.001$, adjusted $\alpha = 0.007$), but more fluctuation (higher entropy score) for boredom ($\chi^2(3) = 26.999$, $p < 0.001$, adjusted $\alpha = 0.040$), confusion ($\chi^2(3) = 29.759$, $p < 0.001$, adjusted $\alpha = 0.033$), correctness ($\chi^2(3) = 185.310$, $p < 0.001$, adjusted $\alpha = 0.010$), and knowledge ($\chi^2(3) = 639.111$, $p < 0.001$, adjusted $\alpha = 0.003$). [We used one-way ANOVA (F-test) for features with equal group variances, and Kruskal-Wallis test (χ^2 test) for features with unequal group variances.]

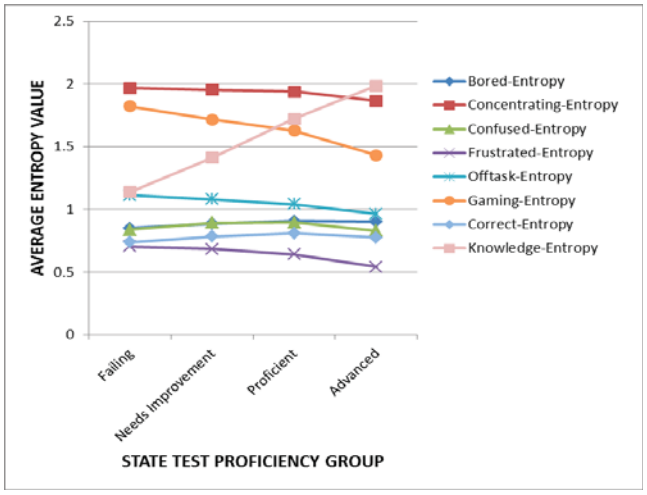

Figure 3. Entropy Scores by MCAS Test Score Category.

These trends suggest that students who performed better in MCAS showed overall consistency across time in exhibiting engaged concentration, frustration, off-task behaviors, and gaming the system, and an overall higher degree of variability across time in exhibiting boredom, confusion, correctness, and knowledge. It is possible that highly successful students may be more aware of their engaged concentration, frustration, off-task, and gaming behaviors within the system, compared to their awareness of the other constructs. Indeed, students who have achieved a higher level of proficiency or mastery of the material may also be more

efficient at controlling and maintaining the negative learning behaviors, and be more engaged. Interestingly, successful students show more variability, indicative of less control, in their boredom, confusion, correctness, and knowledge, possibly due to the nature of the learning task. These successful students may find some problems within ASSISTments too easy or too difficult with respect to their skills, causing them to experience varying degrees of boredom and confusion across time. In other words, the environment may be a major driver of the variability in these constructs. Another possibility comes from results in [24], where more successful students were more likely to be bored or confused when answering original problems, and less bored and confused when answering scaffolding problems. These successful students may also be overconfident in answering problems and become careless [32], exhibiting varying degrees of correctness and knowledge across time.

These relationships suggest that students with higher year-end exam scores were able to control their engagement by becoming less off-task and more consistent in overcoming their frustration and avoiding gaming the system, and be more engaged during their time in ASSISTments. However, a relevant area of future work may be to investigate whether the fluctuations across time for our interaction features are more a function of students' individual differences (e.g. proficiency) and their ability to control their learning behaviors [38], or a function of the learning task (e.g. type of problem, difficulty, etc.) and the learning behaviors it elicits from the students.

While Figure 3 shows the intensity or strength of fluctuations of our constructs across the entirety of student usage of ASSISTments, it does not demonstrate behavior of these fluctuations in fine-grained moments (i.e., persistence or anti-persistence of these constructs; how rapid were the fluctuations?). This is where looking at the Hurst measures of our constructs comes in useful. Figure 4 shows the graph of test proficiency and Hurst measures, where students who have higher MCAS scores achieved lower Hurst scores for engaged concentration ($\chi^2(3) = 134.719$, $p< 0.001$, adjusted $\alpha = 0.017$), frustration ($F(3,1372)= 27.543$, $p< 0.001$, adjusted $\alpha = 0.020$), off-task behavior ($\chi^2(3) = 70.736$, $p< 0.001$, adjusted $\alpha = 0.027$), and confusion ($F(3,1372)= 9.969$, $p< 0.001$, adjusted $\alpha = 0.037$), while higher Hurst scores for knowledge ($\chi^2(3) = 23.935$, $p< 0.001$, adjusted $\alpha = 0.043$) and gaming the system ($\chi^2(3) = 12.425$, $p= 0.006$, adjusted $\alpha = 0.047$).

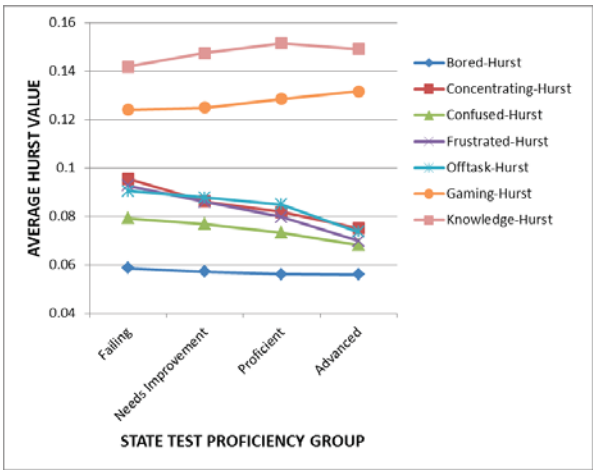

Figure 4. Hurst Scores by MCAS Test Score Category.

This trend in Hurst scores suggests that students who scored high on the MCAS had greater tendency to vary their behaviors, indicative of their actively adapting their learning behaviors. They instead showed regulation strategies in their ability to bounce back from frustration, resolve their confusion, and to re-engage after going off-task. Interestingly, more successful students show more mean reversion in engaged concentration than less successful students. Thus, more successful students were more variable in their engaged concentration (higher probability of concentration at one moment, lower probability of concentration on the next). Along with the Hurst scores for confusion, off-task and frustration, this Hurst trend for engaged concentration may indicate that students who began to feel confused or frustrated switched their focus and went off-task. Conversely, the trend for more successful students showed less variability in their display of knowledge and gaming the system behavior, which would suggest their ability to maintain their high level of knowledge and to not game the system. An understanding of the differences of rate of momentary fluctuations provides a lens on how students who vary in proficiency are able to effectively manage and adjust their affect, behavior, and knowledge within a learning task. It suggests that in the case of ASSISTments, it may be beneficial to teach less successful students strategies for quickly bouncing back from being off-task or ways to resolve their confusion and frustration.

We examine the significance of these differences in trends further by looking at the Pearson correlations between MCAS test scores and student Hurst and Entropy scores for affect, behavior, knowledge, and correctness (Table 1). We also utilize the Benjamini and Hochberg false discovery rate post-hoc correction to adjust the required alpha for significance and to reduce the occurrence of false positives, controlling for inflation of Type 1 error [8].

Table 1. Correlations with MCAS State Test Scores (- significant, p < 0.01; * - significant, p < 0.05)**

Hurst and Entropy Features	r	p-value	Adjusted α
Knowledge-Entropy	.705**	<0.001	0.003
Gaming-Entropy	-.441**	<0.001	0.007
Concentrating- Hurst	-.324**	<0.001	0.010
Frustration-Entropy	-.314**	<0.001	0.013
Correctness-Entropy	.275**	<0.001	0.017
Frustration- Hurst	-.252**	<0.001	0.020
Off-task-Entropy	-.211**	<0.001	0.023
Concentrating-Entropy	-.206**	<0.001	0.027
Off-task- Hurst	-.183**	<0.001	0.030
Confusion- Hurst	-.160**	<0.001	0.033
Bored-Entropy	.139**	<0.001	0.037
Bored-Hurst	-.100**	<0.001	0.040
Knowledge- Hurst	.076**	0.005	0.043
Confusion-Entropy	.076**	0.005	0.047
Gaming- Hurst	.059*	0.029	0.050
Correctness-Hurst	N/A	N/A	N/A

Table 1 shows that there are statistically significant, and reasonably strong relations between MCAS performance and Entropy measures of boredom, engaged concentration, confusion, frustration, off-task, gaming behavior, knowledge and correctness, and Hurst measures of boredom, engaged concentration, confusion, frustration, off-task, gaming behavior and knowledge. Note that for correctness only an Entropy score was calculated as it was a dichotomous measure of a student's answer (1 – correct, 0 – incorrect), and Hurst was not calculated for correctness as Hurst becomes less accurate when the inputs in the time series are discrete rather than continuous (which our other features are).

3.2 Prediction of State Test Scores

To examine the relations of these dynamical measures to MCAS performance, we conducted regression analyses to evaluate the predictive power of these measures (Table 2).

Table 2. State Test Score Model Performance Values Using Different Feature Sets (feature count is after feature selection)

Feature Set	R	R^2	RMSE	Number of Features
Hurst Features Only	0.400	0.160	11.251	5
Entropy Features Only	0.762	0.581	7.941	6
Both Hurst and Entropy Features	0.768	0.590	7.862	9

Combined, Hurst and Entropy assessments of affect, behavior, knowledge and correctness within ASSISTments are predictive of long-term performance (end-of-year state test score, MCAS) with reasonably high model performance. This finding shows that when our automated detectors of affect, behavior, and knowledge are applied at scale, the patterns generated are significantly related to learning outcomes. The specific patterns and contexts in which these interactions occur, however, remain to be further analyzed - for example using methods such as sequential pattern mining or recurrence analysis. Moreover, it is also worth noting that despite the interesting findings discussed above, the model created from dynamical assessments of machine-learned measures of interaction is not much better than a model created from just averaging our interaction features per student (for our sample, this model had a cross-validated R = 0.764) [24]. This suggests that averaging remains a good tool for predicting standardized exam scores, though it does not shed as much light on the phenomena of interest compared to the approach discussed here.

Optimized for predictor significance and model performance, our final model (Table 3) consists of either Hurst or Entropy scores (or both) of boredom, engaged concentration, confusion, frustration, gaming the system, knowledge, and correctness being predictive of MCAS performance.

Our final model leverages the relationships between MCAS and Hurst and entropy measures previously found. Stronger fluctuations across time for knowledge and correctness (positive coefficient for Entropy), and less persistence or quicker reversions in knowledge and engaged concentration (negative coefficient for Hurst), are associated with higher test scores for students. Furthermore, weaker fluctuations across time for boredom, confusion, gaming the system, and frustration (negative coefficient for Entropy), and more persistence or slow fluctuations for gaming the system (positive coefficient for Hurst), are associated with higher test scores for students. These relationships suggest that students with higher year-end exam scores were able to control their engagement by resolving their confusion, bouncing back from being bored, overcoming their frustration, and to show active learning, and be more consistent in not gaming the system during their time in ASSISTments.

Table 3. Final Model of Hurst and Entropy Scores Predicting State Test Scores

Predictors	B	Std. Error	t	Sig
(Constant)	28.821	3.258	8.845	<0.001
Correctness-Entropy	39.566	4.672	8.469	<0.001
Concentrating-Hurst	-34.185	10.738	-3.183	0.001
Gaming-Hurst	22.952	6.853	3.349	0.001
Knowledge-Hurst	-22.935	4.579	-5.009	<0.001
Bored-Entropy	-21.318	2.773	-7.687	<0.001
Frustration-Entropy	-17.874	1.892	-9.447	<0.001
Knowledge-Entropy	17.463	0.723	24.169	<0.001
Gaming-Entropy	-9.371	1.126	-8.320	<0.001
Confusion-Entropy	-6.157	1.803	-3.416	0.001

4. DISCUSSION AND CONCLUSION

In this paper, we utilized dynamical methodologies to investigate how nuanced patterns of affect, behavior, knowledge, and correctness were related to and predictive of students' end-of-year exam scores. Fine-grained models of student affect (boredom, engaged concentration, confusion, frustration) behavior (off-task behavior, gaming the system), and knowledge were applied to data from 1,376 students who used an educational software in mathematics over the course of a year during their middle school to generate interaction features. We then utilized dynamical measures of Hurst exponents and Entropy analysis to quantify the degree of randomness (or non-randomness) present within patterns of these interaction patterns.

Our results show that these dynamical assessments of students' interactions throughout the year (affect, behavior, knowledge, and correctness) are significantly associated with their end-of-year performance in a state test. Entropy scores of students for all of our interaction features showed significant differences between students in varied test proficiencies (as measured by the year-end exam). Across time, the more control a student demonstrated in frustration, engaged concentration, off-task behaviors, and gaming the system behaviors, as well as more flexibility in boredom, confusion, knowledge and correctness, the higher the student scored on the year-end exam. Students' Hurst scores also showed significant relations with the learning outcome, where students with more occurrences of fluctuations for engaged concentration, confusion, frustration, and off-task behaviors, and more persistence for knowledge and gaming the system were likely to perform better. These relations were supported by these dynamical assessments being predictive of performance in the end-of-year state test.

It is notable that most Hurst exponent values fell well below 0.5, indicating that overall, fine-grained machine-learned estimates of affect, behavior, knowledge in the system interaction of the 1,376 students are not random, and according to students' state or the learning task within the system, students show signs of switching between various degrees of affect, behavior, and knowledge over

time. In the future, it may be useful to examine sequential patterns of each interaction feature, looking also at the context and circumstances in the usage of the system that lead to students having increasing or decreasing occurrences (as well as points of inflection) in affect, behavior, and knowledge. The Hurst and Entropy may be able to be used in real-time to capture these affective changes and then provide feedback to a user model (or teacher) about the student. Less successful students may be made aware of their learning behaviors so they may more effectively regulate them, in particular for frustration, confusion, off-task-behavior, and gaming the system. They may also be taught strategies to more quickly bounce back from being off-task or even resolve their frustration and confusion.

Overall, these exploratory findings obtained when we dynamically assess the measures of interaction take a step further in evaluating how fine-grained machine-learned assessments of affect, behavior, and knowledge relate to learning outcomes. Looking at patterns using a combination of machine-learning techniques provides an avenue for observing the degree to which students regulate their actions in a learning task. Self-regulation research shows that when students are motivated to achieve learning goals they are more likely to regulate their behaviors [7]. This current study provides a preliminary lens on how dynamic measures of fine-grained series of distinctive affect (academic emotions) and behavior (engagement) are reflective of students' emotional and motivational regulation within a learning environment [9, 18], as well as the roles of affect and behavior on self-regulated learning [25].

5. ACKNOWLEDGMENTS

This research was supported by grants NSF #DRL-1031398, NSF #SBE-0836012, grant #OPP1048577 from the Bill and Melinda Gates Foundation, and grant #R305A130124 from the Institute of Education Sciences.

6. REFERENCES

[1] Baker, R.S.J.d. 2007. Modeling and Understanding Students' Off-Task Behavior in Intelligent Tutoring Systems. In *Proceedings of ACM CHI 2007: Computer-Human Interaction*, 1059-1068.

[2] Baker R.S.J.d., Corbett A.T., Gowda S.M., Wagner A.Z., MacLaren B.M., Kauffman L.R., Mitchell A.P., and Giguere S. 2010. Contextual Slip and Prediction of Student Performance after Use of an Intelligent Tutor. In *Proc. UMAP 2010*, 52-63.

[3] Baker, R.S.J.d., Gowda, S.M., Wixon, M., Kalka, J., Wagner, A.Z., Salvi, A., Aleven, V., Kusbit, G., Ocumpaugh, J., and Rossi, L. 2012. Towards Sensor-free Affect Detection in Cognitive Tutor Algebra. In *Proceedings of the 5th International Conference on Educational Data Mining*, 126-133.

[4] Baker, R.S., Corbett, A.T., and Koedinger, K.R. 2004. Detecting Student Misuse of Intelligent Tutoring Systems. In *Proceedings of the 7th International Conference on Intelligent Tutoring Systems*, 531-540.

[5] Baker, R.S.J.d., Corbett, A.T., Koedinger, K.R., Evenson, S.E., Roll, I., Wagner, A.Z., Naim, M., Raspat, J., Baker, D.J., and Beck, J. 2006. Adapting to When Students Game an Intelligent Tutoring System. In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, 392-401.

[6] Baker, R. S., D'Mello, S. K., Rodrigo, M. M. T., and Graesser, A. C. 2010. Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, 68(4), 223-241.

[7] Bandura, A. 1991. Social cognitive theory of self-regulation. *Organizational behavior and human decision processes,* 50, 2, 248-287.

[8] Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 289-300.

[9] Bosch, Nigel, and Sidney D'Mello. 2013. Sequential Patterns of Affective States of Novice Programmers. *The First Workshop on AI-supported Education for Computer Science (AIEDCS 2013)*.

[10] Cocea, M., Hershkovitz, A., and Baker, R.S.J.d. 2009. The Impact of Off-task and Gaming Behaviors on Learning: Immediate or Aggregate? In *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, 507-514.

[11] Conati, C., and Maclaren, H. 2009. Empirically building and evaluating a probabilistic model of user affect. *User Modeling and User-Adapted Interaction,* 19, 3, 267-303.

[12] Corbett, A.T., and Anderson, J.R. 1995. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction,* 4, 4, 253-278.

[13] D'Mello, S.K., Craig, S.D., Witherspoon, A. W., McDaniel, B. T., and Graesser, A. C. 2008. Automatic Detection of Learner's Affect from Conversational Cues. *User Modeling and User- Adapted Interaction*, 18 (1-2), 45-80.

[14] D'Mello, S. K. and Graesser, A. C. 2012. Dynamics of Affective States during Complex Learning. *Learning and Instruction*, 22, 145-157.

[15] Fasolo, B., Hertwig, R., Huber, M., and Ludwig, M. 2009. Size, entropy, and density: What is the difference that makes the difference between small and large real-world assortments? *Psychology & Marketing*, 26, 3, 254-279.

[16] Feng, M., Heffernan, N.T., and Koedinger, K.R. 2009. Addressing the assessment challenge in an online system that tutors as it assesses. User Modeling and User-Adapted Interaction: *The Journal of Personalization Research (UMUAI)*, 19, 3, 243-266.

[17] Grossman, E. R. F. W. 1953. Entropy and choice time: The effect of frequency unbalance on choice-response. Quarterly *Journal of Experimental Psychology,* 41-51.

[18] Gumora, G. and Arsenio, W. F. 2002. Emotionality, emotion regulation, and school performance in middle school children. *Journal of School Psychology*, 40, 5, 395-413.

[19] Hadwin, A. F., Nesbit, J. C., Jamieson-Noel, D., Code, J., and Winne, P. H. 2007. Examining trace data to explore self-regulated learning. *Metaknowledge and Learning*, 2, 107–124.

[20] Hershkovitz, A., Baker, R.S.J.d., Gobert, J., Wixon, M., and Sao Pedro, M. 2013. Discovery with Models: A Case Study on Carelessness in Computer-based Science Inquiry. *American Behavioral Scientist*, 57, 10, 1479-1498.

[21] Lee, D. M. C., Rodrigo, M. M. T., d Baker, R. S., Sugay, J. O., and Coronel, A. 2011. Exploring the relationship between novice programmer confusion and achievement. In *Proceedings of Affective Computing and Intelligent Interaction*, 175-184.

[22] Liu, Z., Ocumpaugh, J., and Baker, R. S. 2013. Sequences of Frustration and Confusion, and Learning. In *Proc. Int. Conf. Ed. Data Mining*, 114-120.

[23] Ocumpaugh, J., Baker, R., Gowda, S., Heffernan, N., and Heffernan, C. 2014. Population validity for Educational Data Mining models: A case study in affect detection. *British Journal of Educational Technology*, 45, 3, 487-501.

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

[25] Pekrun, R., Goetz, T., Titz, W., and Perry, R. P. 2002. Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational Psychologist*, 37, 2, 91-105.

[26] Peng, C. K., Buldyrev, S. V., Havlin, S., Simons, H. E., Stanley, H. E., and Goldberger, A. L. 1994. Mosaic organization of DNA nucleotides. *Physical Review E*, 49, 1685-1689.

[27] Razzaq, L., Feng, M., Nuzzo-Jones, G., Heffernan, N.T., Koedinger, K. R., Junker, B., Ritter, S., Knight, A., Aniszczyk, C., Choksey, S., Livak, T., Mercado, E., Turner, T.E., Upalekar. R, Walonoski, J.A., Macasek. M.A. and Rasmussen, K.P. 2005. The Assistment project: Blending assessment and assisting. In *Proc. AIED 2005*, 555-562.

[28] Ritter, S., Joshi, A., Fancsali, S. E., and Nixon, T. 2013. Predicting Standardized Test Scores from Cognitive Tutor Interactions. In *Proceedings of the 6th International Conference on Educational Data Mining*, 169-176.

[29] Rodrigo, M. M. T., Baker, R. S., and Nabos, J. Q. 2010. The relationships between sequences of affective states and learner achievement. In *Proceedings of the 18th International Conference on Computers in Education*, 56-60.

[30] Sabourin, J., Mott, B., and Lester, J. 2011. Modeling Learner Affect with Theoretically Grounded Dynamic Bayesian Networks. In *Proc. ACII 2011*, 286-295.

[31] San Pedro, M.O.Z., Baker, R.S.J.d., Bowers, A.J., Heffernan, N.T. 2013. Predicting College Enrollment from Student Interaction with an Intelligent Tutoring System in Middle School. In *Proceedings of the 6th International Conference on Educational Data Mining*, 177-184.

[32] San Pedro, M.O.Z., Baker, R.S.J.d., Rodrigo, Mercedes, M.M.T. 2014. Carelessness and Affect in an Intelligent Tutoring System for Mathematics. *International Journal of Artifiical Intelligence in Education*, 24, 189-210.

[33] San Pedro, M.O.Z., Ocumpaugh, J.L., Baker, R.S., Heffernan, N.T. 2014. Predicting STEM and Non-STEM College Major Enrollment from Middle School Interaction with Mathematics Educational Software. In *Proceedings of the 7th International Conference on Educational Data Mining*, 276-279.

[34] Shannon, C. 1951. Prediction and Entropy of printed English. *Bell Systems Technical Journal*, 27, 50-64.

[35] Snow, E. L., Allen, L. K., Jacovina, M. E., and McNamara, D. S. 2015. Does agency matter?: Exploring the impact of controlled behaviors within a game-based environment. *Computers and Education,* 26, (2015), 378-392.

[36] Snow, E. L., Allen L. K., Russell, D. G., and McNamara, D. S. 2014. Who's in control?: Categorizing nuanced patterns of behaviors within a game-based intelligent tutoring system. In J. Stamper, Z. Pardos, M. Mavrikis, and B. M. McLaren (Eds.), Proce*edings of the 7th International Conference on Educational Data Mining* (London, UK July 4 -7, 2014). Heidelberg, Berlin, Germany: Springer, pp. 185-192.

[37] Snow, E. L., Jackson, G. T., and McNamara, D. S. 2014. Emergent behaviors in computer-based learning environments: Computational signals of catching up. *Computers in Human Behavior,* 41, (2014), 62-70.

[38] Snow, E. L., Jacovina, M. E., Allen, L. K., Dai, J., and McNamara, D. S. 2014. Entropy: A stealth assessment of agency in learning environments. In J. Stamper, Z. Pardos, M. Mavrikis, B. M. McLaren (eds.), Proce*edings of the 7th International Conference on Educational Data Mining* (London, UK July 4 -7, 2014). Heidelberg, Berlin, Germany: Springer, pp. 185-192.