Studying Affect Dynamics and Chronometry Using Sensor-Free Detectors

Anthony F. Botelho
Worcester Polytechnic Institute
100 Institute Rd
Worcester, MA
abotelho@wpi.edu

Ryan S. Baker
University of Pennsylvania
3700 Walnut St
Philadelphia, PA
ryanshaunbaker@gmail.com

Jaclyn Ocumpaugh
University of Pennsylvania
3700 Walnut St
Philadelphia, PA
jlocumpaugh@gmail.com

Neil T. Heffernan
Worcester Polytechnic Institute
100 Institute Rd
Worcester, MA
nth@wpi.edu

ABSTRACT
Student affect has been found to correlate with short- and long-term learning outcomes, including college attendance as well as interest and involvement in Science, Technology, Engineering, and Mathematics (STEM) careers. However, there still remain significant questions about the processes by which affect shifts and develops during the learning process. Much of this research can be split into affect dynamics, the study of the temporal transitions between affective states, and affective chronometry, the study of how an affect state emerges and dissipates over time. Thus far, these affective processes have been primarily studied using field observations, sensors, or student self-report measures; however, these approaches can be coarse, and obtaining finer-grained data produces challenges to data fidelity. Recent developments in sensor-free detectors of student affect, utilizing only the data from student interactions with a computer-based learning platform, open an opportunity to study affect dynamics and chronometry at moment-to-moment levels of granularity. This work presents a novel approach, applying sensor-free detectors to study these two prominent problems in affective research.

Keywords
Student Affect, Affect Dynamics, Affect Chronometry, Deep Learning, Sensor-Free Detectors

1. INTRODUCTION
The various affective states experienced by students during learning have received significant attention from the research community for their prominence in the learning process. Student affect has been shown to correlate with several measures of student achievement [6][22][28], has been found to be predictive of whether students attend college several years later [24], and also whether students choose to take steps towards careers in Science, Technology, Engineering, and Mathematics (STEM) fields [30]. While significant steps have been taken toward understanding the interrelationships between of affect and learning, there are many questions that remain unanswered with regard to how affect is exhibited by students over time as well has how such temporal trends may be informative of student learning outcomes.

The temporality of student affect has been characterized into two areas of study, affect dynamics [31] and affective chronometry. Affect dynamics studies temporal shifts in affect to understand which transitions between affective states are most common. A theoretically-grounded model of affective dynamics has been proposed by D'Mello and Graesser [10], which suggests a typical resolution cycle, where students transition from engaged concentration to surprise to confusion and back to engaged concentration, but which also hypothesizes alternative transitions, including a path from confusion to frustration and boredom.

Affective chronometry also uses temporal measures, but focuses more closely upon how individual affective states (e.g., boredom) behave over time. This was first studied as a special case of affective dynamics, where researchers investigated how frequent it was for an affective state to transition to itself (aka "self-transitions"). More recently, D'Mello and Graesser [9] proposed instead investigating an affective state's "half life," or the decay in the probability of an affective state persisting for a specific duration of time. [9] found evidence that six affective states exhibit exponential decay in their probability over time. That is, the probability that a student remains in a particular state decreases exponentially as the amount of time that the student persists in that state increases. However, engaged concentration (referred to as flow) showed a much slower decay rate than other affective states (e.g., frustration).

There is now a growing body of research in affective dynamics and affective chronometry, commonly using field observations [26][13], or self-reports accompanied by video data [3][9]. These important studies have helped to advance the field, but each method imposes different kinds of limitations on the grain-size of the data. Continuous observation is impractical both for self-report and field observation studies, and it is highly time-consuming for video recording (which can also break down when the student moves away from his or her desk, either for off-task reasons or for on-task purposes like peer-tutoring or requesting assistance). Despite the limitations of these methods, they have often been preferred to sensor-free detectors of affect due to higher reliability/quality of the data obtained. However, recent advances in sensor-free detection of affect, based on deep learning methods, have substantially increased the quality of models [5], making interaction-based detectors a viable alternative. While these models are also not without limitations, their improved performance provides an alternative that facilitates near-continuous labeling at scale. As such, the recent advent of higher-quality detectors introduce the opportunity to study affect dynamics and affective chronometry with fine levels of granularity at scale.

In this paper, we present research studying affect dynamics and affective chronometry with the use of deep learning sensor-free affect detectors. We report the affect dynamics and chronometry for four commonly-studied affective states: engaged concentration [7] (also referred to as engagement, flow, and equilibrium), boredom [7][19], confusion [6][16], and frustration [16][23]. We investigate these relationships in the real-world learning of just under a thousand students, and compare our findings to prominent foundational research [9][10].

2. PREVIOUS WORK

The theoretical model of affective dynamics proposed by D'Mello and Graesser [10] has become widely recognized in the study of affective state transitions. The model proposes a set of theoretically hypothesized transitions that have emerged through the study of student affect, as illustrated by the simplified representation of the model in Figure 1. While the full model observes numerous affective states including surprise and delight, we restrict the analysis in this paper to the key affective states of engaged concentration, boredom, confusion, and frustration.

The model hypothesizes that specific transitions between affective states are particularly common. In this model, a student commonly begins in a state of equilibrium (i.e. flow or engaged concentration). The student remains in this state until novelty or difficulty emerges, at which point the student may transition to confusion. The student may transition back to engaged concentration by resolving this confusion, possibly experiencing delight upon the way. Alternatively, the student my transition from confusion to frustration, at which point the model suggests that the student is unlikely to transition back to the more productive cycle of engaged concentration and confusion; instead, the student is more likely to transition from frustration to boredom. As such, while students may be expected to oscillate between certain adjacent states in the model, the model suggests that it is unlikely for students to transition to unconnected states as depicted in Figure 1.

The model has been explored in several studies [27][8] observing differences in student affect, and has become influential to other research studying affect dynamics in the context of other constructs such as gaming the system [26]. Other studies prior to the publication of this model also studied affective dynamics [1][29]. While the specific affective states studied across these projects vary, the four affective states studied in this work are among the most commonly observed in this area of research. However, work in other paradigms also exists; for example, Redondo [25] attempted to identify when a student's affect shifts from increasingly positive to becoming more negative, or vice-versa, in self-report Likert scale data, finding that unexpectedly positive or negative affect typically indicated a shift in overall affective trajectory. However, she did not compare the prevalence of turning points found to overall base rates of affect, or analyze the chronometry of the sequences she studied. In general, across these papers, estimates of student affect have been collected through a range of methodologies including, most commonly, quantitative field observations (QFOs) [13][12][26][20], but also through self-reports in conjunction with post-hoc judgements of recorded video [3][4].

While there have been a large number of projects investigating affective dynamics, there has been substantially less research pertaining to affective chronometry. The study of affective chronometry is at times seen in affective dynamics papers. Among the papers investigating affective dynamics, several studies, including that of Baker, Rodrigo, and Xolocotzin [1] have found that state self-transitions, where the student is in the same affective state in one observation as in the previous observation, were often statistically significantly more likely than chance. This suggests that students in each state do tend to persist for at least the duration of the time interval between observations (1 minute in that article); however, this paper did not observe the chronometry beyond this interval. In foundational work in this area, D'Mello and Graesser [9] investigated the duration of different affective states, proposing a methodology with which to evaluate the "half-life," or decay of individual affective states experienced by students. Using a computer-based system known as AutoTutor, the authors used a combination of self-reports of the students and expert and peer judgments of student affect made using recorded video in order to measure and evaluate the length of time students commonly remained in each experienced affective state. However, that work was conducted on a relatively small number of subjects working on AutoTutor in a lab setting, on a task not related to their studies. It is therefore unclear whether the findings obtained in that context will generalize to data from a classroom environment where students are working on authentic educational tasks. The same methodology for measurement and evaluation of affective chronometry as presented in that work will be applied here to understand and compare affective chronometry – however, instead of using self-report, this project will utilize sensor-free detectors of affect applied to data collected from real students working in classroom environments.

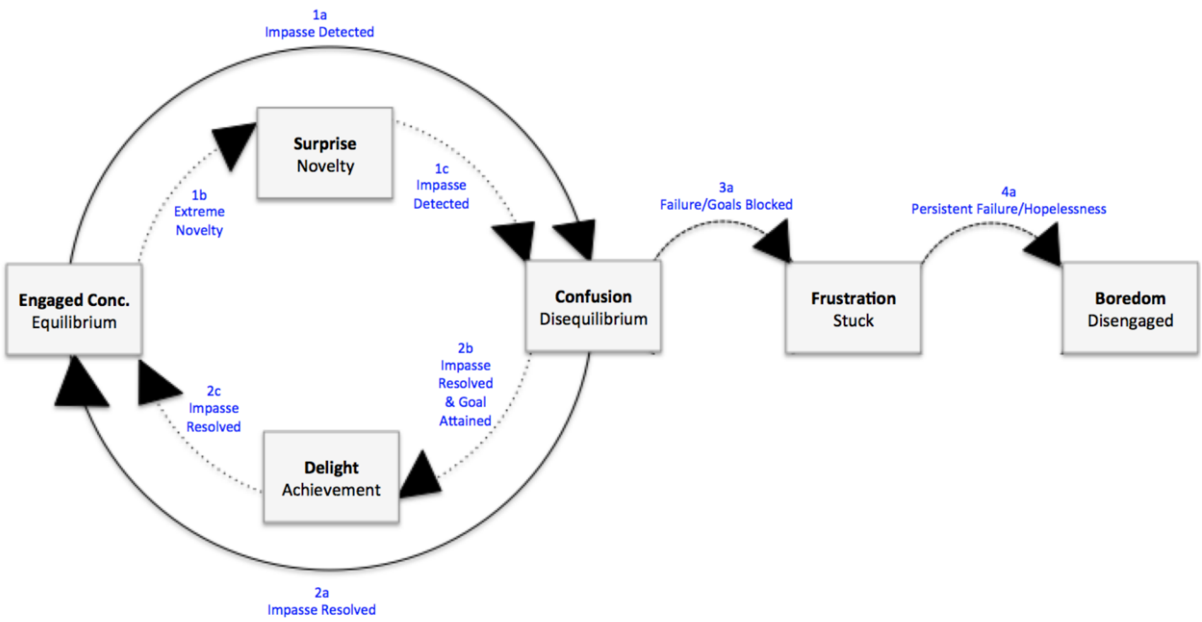

Figure 1: The proposed theoretical model of affect dynamics as presented by D'Mello and Graesser [10]

2.1 Detectors of Student Affect

We apply the sensor-free detectors of student affect previously described in Botelho et al. [5] to our data in order to study affective dynamics and chronometry. We use the same data set in this work from which the training set originally used in Botelho et al. [5] was sampled, to ensure maximum validity of the detectors. In applying the detectors to this data set, we determined that several minor adjustments needed to be made to the detectors, so that the training data set was aligned to the ground truth observations in a way that could be more easily applied to the unlabeled data. We also reduced the number of features used as input to the model building algorithm. The detectors were refit using this adjusted dataset and produced performance metrics comparable to the previous work (average AUC = .74, average Cohen's Kappa = 0.20).

As in Botelho et al. [5], these sensor-free detectors were developed using a long short term memory (LSTM) [15] network, a type of deep learning model designed for time series data. LSTM networks use a large number of learned parameters with internal memory that can model temporal trends within the data to make estimates that are better informed by previous time steps within the series. Although the initial training sample was imbalanced, the use of resampling did not improve model performance, and a min-max estimate scaling was used instead. The LSTM model is trained as a sequence-to-sequence model, meaning that it accepts an entire sequence of time steps as input and produces a sequence of outputs. These outputs are in the form of a sequence of estimates of the probability that each of four affective states of engaged concentration, boredom, confusion, and frustration are occurring at each 20-second time step, or "clip," within the data. We use this sequence of probabilities to study affective dynamics and chronometry – the details of these analyses are provided in later sections. The LSTM model was found to produce cross-validated AUC values that substantially outperformed prior sensor-free detectors, which had previously exhibited an average AUC = 0.66, developed using older algorithms with the same dataset [21][32]. In addition, LSTM models are designed to exploit the temporal character of the data, suggesting that they will be able to model temporal changes and transitions between affective state better than a model that treats each 20-second clip of student behavior as an independent sample.

3. METHODOLOGY

3.1 Dataset

The data[1] used in this work is comprised of action-level student data collected within the ASSISTments learning platform [14]. ASSISTments is a computer-based learning system used daily by thousands of students in real classrooms (over 50,000 a year) and hosts primarily middle school math content. The system has been used in several previous papers to study student affect, in many cases using sensor-free detectors of student affect.

Within this paper, we utilize a dataset originally used to develop sensor-free automated detectors of student affect. Detectors were originally developed using data collected by conducting field observations of student affect as 838 students used ASSISTments. 3,127 20-second field observations were collected in total, with gaps between one and several

[1] The data used in this work is made available at http://tiny.cc/EDM2018_affectdata

minutes between observations of the same student. For this paper, we analyze the entire data set of interaction for those 838 students on the days when observation occurred, 48,276 20-second segments of student behavior in total. We format the data in terms of 20-second segments of behavior in order to use the sensor-free detectors of affect, which were developed at this grain size (in line with the original field observations, which were conducted at the same grain size). The original training data set was highly imbalanced, with approximately 82% of observations coded as engaged concentration, 10% coded as boredom, 4% coded as confused, and 4% coded as frustration. This imbalance is consistent with previous research on the prevalence of these affective categories in systems such as ASSISTments.

The sensor-free LSTM detectors were applied to this dataset, providing an estimate of the probability of each of the four observed affective states for each of the 20-second segments of behavior within the system. The ground-truth labels used in model training are removed from this dataset and instead are replaced with the estimates produced by the sensor-free detectors. We replaced the ground-truth labels with the detector outputs so that the data would be comparable across all of the 48,276 observations.

3.2 Affect Dynamics

The estimates produced by the sensor-free detectors, when applied to the analysis dataset, are used to observe which transitions between affective states are frequent and statistically significantly more likely than chance. As is described in the previous section, the model produces four continuous-valued estimates corresponding with the 4 affective states of engaged concentration, boredom, confusion, and frustration. However, these estimates must be discretized and reduced to a single label describing the most likely affective state exhibited by the student at each time step. It is not sufficient to simply conclude that the most probable affective state (e.g. the affective state with the highest confidence) is the current affective state. For example, the model may predict very small values for all four affective states.

Instead, we first select a threshold that indicates that a specific affective state is likely occurring during a specific clip. We use a threshold of 0.5, defining a value above this threshold to be indicative of the presence of that corresponding affective state for the time step. 0.5 is a reasonable threshold as the detectors were previously run through a min-max scaling of the model outputs to remove majority class bias (cf. [5]). However, there exists the possibility, as expressed in the example above, that no estimate across the four affective states surpasses this defined threshold. In such cases, a fifth "Neutral/Other" affective state is introduced to represent that none of the affective states we are studying is occurring; this state has been included in similar previous analyses of affect dynamics as well ([13][12][29][27][4][9]). Conversely, it is possible for more than one estimate across the four outputs to surpass the defined threshold. In this unusual case (less than 1% of our data), no single affective state label can be applied and this clip (and transitions from and to this clip) is omitted from the subsequent analyses.

Once all estimates have been classified as either a single affective state or the neutral state, transitions between these states within each student are computed. As in [10], we omit self-transitions where the student remains in their current affective state; these are instead represented through affective chronometry (see next section). We report D'Mello's L [11] as a measure of the commonality of each possible transition from a source affective state to a destination affective state along with a corresponding p-value denoting the probability of this frequency of transition being obtained by chance. The D'Mello's L metric can be interpreted in a similar manner to Cohen's kappa, describing the degree to which each transition is more (or less) likely than would be expected according to the overall proportion of occurrence of the destination affective state across all cases. Values of D'Mello's L below zero are less likely than chance; values above zero represent the percent more likely than chance the finding is. In other words, a D'Mello's L of 0.4 represents a transition that occurs 40% more often than would be expected from the destination state's base rate. We compute statistical significance of these transitions using the method originally proposed in [11] – D'Mello's L is computed for each student and transition, and then the set of transitions is compared to 0 using a one-sample two-tailed t-test. Benjamini and Hochberg's [2] correction is used to control for the substantial number of statistical comparisons conducted.

3.3 Affective Chronometry

Our methodology for affective chronometry closely follows that of D'Mello and Graesser [9], with whom we compare our findings. In their analysis, the rate of decay was calculated as a probability of each state persisting over a 60-80 second window, using affect labels aggregated across multiple observation methods including the use of self-reports and both peer- and expert-observers. The probability that each affective state persisted (i.e. $\Pr(\mathrm{E}_t = \mathrm{E}_{t+20})$) was computed for 20 second intervals within that window.

The analysis in this paper uses the same discretized affect labels described in the previous section, transforming a sequence of sets of four probabilities to a single most-likely affective state per clip. The sequence of labels is broken into a set of episodes of each affective state, where an episode describes a series of non-transitioning affect that starts when the student transitions into the state and ends when the student transitions out of the state. A cumulative sum of time, in seconds, is calculated for each episode to measure how long each student remained in each affective state. With this value, a probability that a state will persist beyond a defined number of seconds can be calculated.

Due to the nature of our affect detection approach, persistence is estimated in 20 second intervals. At each interval, the probability that a student remains in eachtheir current affective state is calculated for durations up to 300 seconds, or 5 minutes. The resulting 16 probabilities (for durations of 0, 20, 40, ... , 300 seconds) can then be used to compare the rates of decay across each of the observed affective states.

4. RESULTS

4.1 Observing Affect Dynamics

The affective state transitions, measured by D'Mello's L, are reported in Table 1 with accompanying significance. Aside from those transitions that occur to/from the neutral/other

Table 1: The transitions between affective states. D'Mello's L values are shown. Transitions that are statistically significantly more likely than chance, after Benjamini and Hochberg's post-hoc correction, are denoted *.

From State	To State	D'Mello's L	p-value
Engaged Concentration	Engaged Concentration	—	—
	Boredom	0.260*	<0.001
	Confusion	0.004	0.136
	Frustration	-0.12*	0.012
	Neutral/Other	0.481*	<0.001
Boredom	Engaged Concentration	0.194*	<0.001
	Boredom	—	—
	Confusion	-0.004	0.208
	Frustration	0.036*	<0.001
	Neutral/Other	0.235*	<0.001
Confusion	Engaged Concentration	0.341*	0.006
	Boredom	-0.127*	<0.001
	Confusion	—	—
	Frustration	-0.026*	0.001
	Neutral/Other	-0.156	0.157
Frustration	Engaged Concentration	0.279*	<0.001
	Boredom	-0.107*	<0.001
	Confusion	0.008	0.391
	Frustration	—	—
	Neutral/Other	0.279*	<0.001
Neutral/Other	Engaged Concentration	0.753*	<0.001
	Boredom	-0.057*	<0.001
	Confusion	0.003	0.302
	Frustration	0.015*	0.007
	Neutral/Other	—	—

state, the most common significant transition appears to occur between confusion and engaged concentration, followed by that of frustration to engaged concentration. Contrary to the theoretical model proposed by D'Mello and Graesser [10], significant transitions are found between engaged concentration and boredom as well as from boredom to engaged concentration. The findings suggest that students do not transition between these states through others as in the proposed theoretical model, but can occur directly.

It is further illustrated in the table that no state is found to transition to confusion more likely than chance, for which there are several possible explanations. Confusion was the least-frequently detected state as estimated by the sensor-free model (under 1.0% of the dataset). As such, it is likely that there simply were not enough instances of detected confusion in the data to produce significant results, possibly because the model had difficulty detecting confusion, contributing to an under-sampling of this state as estimated by the model.

These positive and significant transitions as identified by Table 1 are illustrated in Figure 2 for better comparison to the theoretical model depicted in Figure 1. Not only do the already-identified transitions become clearer, the number of transitions occurring to and from the neutral/other state, listed simply as "no label" in that figure, are also made prominent. As described in the generation of this fifth state, this represents those estimates where no model estimates across the four affective states exceeded the defined threshold. It is important to note that this state may not be a single state at all, but rather comprehensively represents all other affective states exhibited by students that are not observed in the analysis. As such, it is difficult to make meaningful claims or draw significant conclusions regarding transitions occurring to or from this state.

The divergence of the emerging transitions and the theoretical model indicate that there are fewer oscillations that are detected by the machine-learned method. While not included in the theoretical model, D'Mello and Graesser propose in the same work [10] that oscillations can occur between all adjacent affective states within the graph under certain conditions, but that is certainly not the case as seen in Figure 2 gained from the empirical results of this work. This suggests that the learned model finds that students do

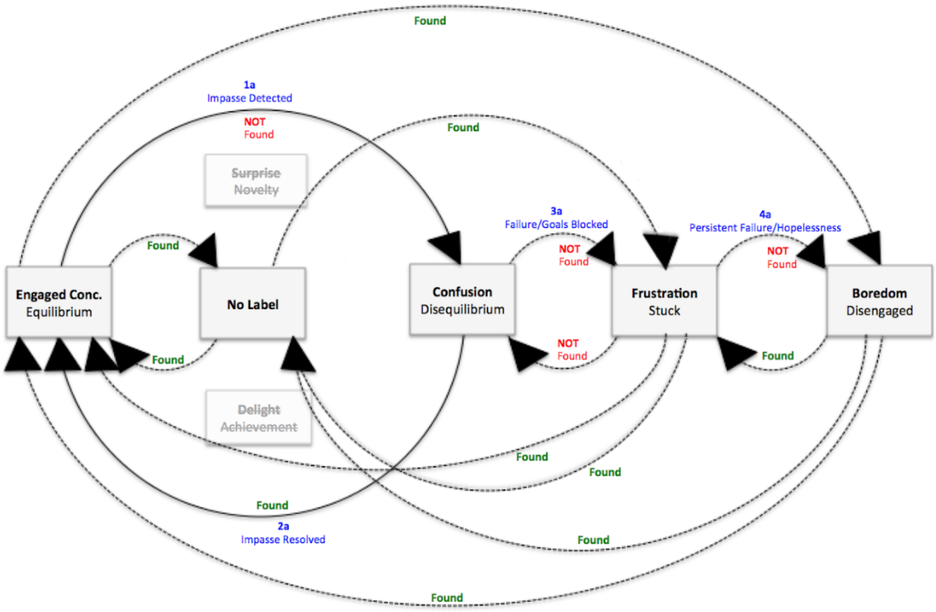

Figure 2: The resulting positive and significant affect transitions as compared to the D'Mello and Graesser [10] theoretical model.

not commonly transition back and forth between states such as confusion and frustration as often as hypothesized by the theoretical model, but no other such cases emerge.

4.2 Observing Affective Chronometry

The results of our affective chronometry analysis illustrate the length of time students commonly spend in each affective state before transitioning to either another observed state or the neutral/other state. The results of this analysis, depicted in Figure 3, show notable differences in affective half-life between affective states. Engaged concentration and boredom exhibit much more gradual declines as opposed to both confusion and frustration which both exhibit steep and rapid decay. Just as was done in the previous work of D'Mello and Graesser [9], the decay can be quantified by fitting an exponential function to each of the observed states. Again, as the neutral/other state may comprehensively represent multiple states that are not measured in this work, this state is not included in the analyses of affective chronometry; if included, the results may simply illustrate an average decay over non-included affective states.

The value of decay for each state, as calculated by fitting an exponential curve to each states probability of persisting (Pr(No Change)) over time. Engaged concentration (decay = -0.003) and boredom (decay = -0.004) are found to have similarly gradual decay as compared to that of the remaining two states. Frustration (decay = -0.01) and confusion (decay = -0.024) are found to decay significantly faster. Of the studied states, only confusion is found to fail to persist past 5 minutes.

While the affective decay of engaged concentration, boredom, and frustration follow the general trend found by the work of D'Mello and Graesser in previous work [9], confusion deviates from this alignment. This difference is illustrated by Figures 4 and 5. Figure 4 illustrates the plotted exponential fit lines that were learned from the estimates produced by the sensor-free detectors. For comparison, Figure 5 illustrates the plotted exponential decay, as reported in Table 1 of D'Mello and Graesser [9]. From this, it becomes apparent that confusion is found to exhibit similar decay patterns to that of engaged concentration and boredom, being more gradual over time, than that of frustration.

The other distinctive difference that emerges from the comparison of Figures 4 and 5 is that of the average time for decay across all affective states. This suggests that the average time that students remain in any affective state, as determined by the sensor-free model, is consistently longer than those found in D'Mello and Graesser [9]. The previous work reports that students rarely remained in a single state for longer than 60 seconds, and, following the learned exponential curve in Figure 5, no state seems to persist beyond 3 minutes, with most states reaching a probability of persisting close to 0 long before that time point. In comparison, each of the affective states, with the exception of confusion, are found to persist past the 5 minute time point, with engaged concentration and boredom seemingly persisting significantly beyond this point. Even in considering the 60 second timeframe, the fastest decaying state of confusion exhibits students persisting beyond this interval.

The divergence of the decay rates as exhibited by the estimates of the sensor-free model and those of the empirical findings reported in [9] may be due to a combination of differences between the two works. One possible explanation is the difference in learning contexts and the different learning interactions being studied in each of the two works. In this work, for example, the students comprising

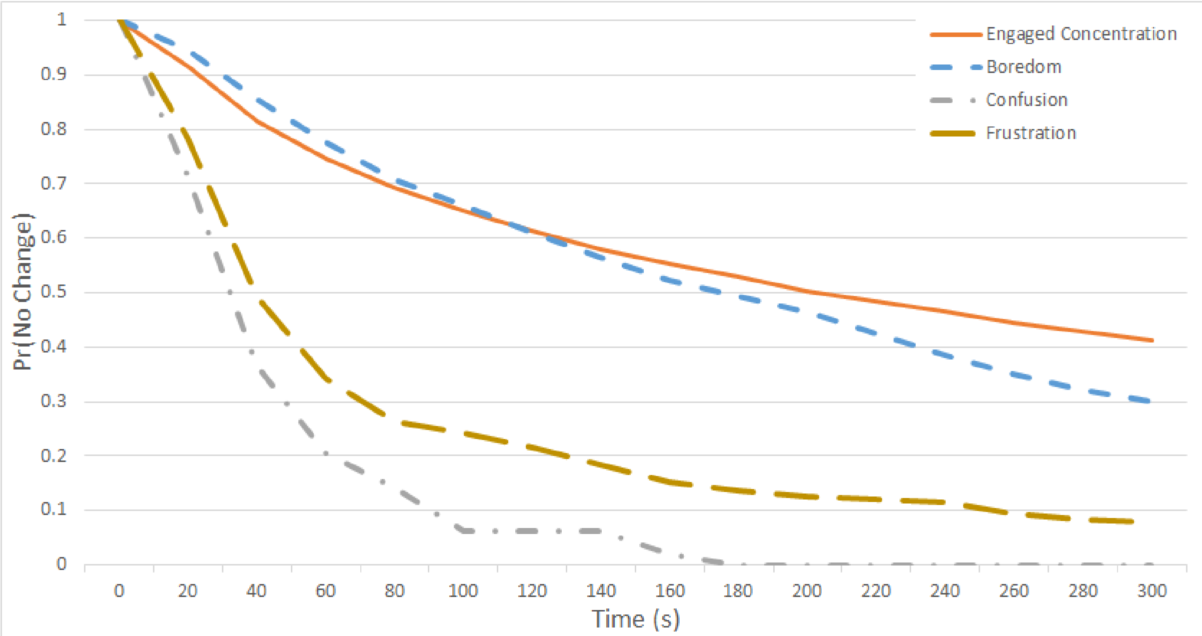

Figure 3: The probability of a student persisting in each affective state over time.

the dataset were in a classroom environment interacting with the computer-based system of ASSISTments. The previous study reported by [9], had students interacting with different software, namely that of AutoTutor, and also took place in a controlled lab setting. The domain of study also exhibits differences in that the students in AutoTutor were answering questions pertaining to computer literacy that are described as requiring students to answer in several sentences. The students using ASSISTments, however, were middle school students working on math content. The differences between both the content and the environment could have a distinct effect on the states of affect exhibited by students as well as the length of time students persist in each affective state.

5. DISCUSSION AND FUTURE WORK

The current work presents, to the knowledge of the authors, the first application of sensor-free affect detectors to study affect dynamics and affective chronometry. In studying affective dynamics, we can compare our results to a past theoretical model of affect dynamics proposed by D'Mello and Graesser [10], as well as other past empirical work. In affective chronometry, we can compare our results to past work [9], also by D'Mello and Graesser. The resulting model of affect dynamics produced by the application of sensor-free detectors shares little with the theorized model in regard to the significant transitions that emerged. Most notably, our model suggests oscillations between engaged concentration and boredom which are hypothesized not to occur significantly in the theorized model; it has been found in other empirical work, however, that transitions between engaged concentration and boredom do appear [3][4]. The model of affective chronometry finds a similar pattern to D'Mello and Graesser in terms of which affective states are shorter and longer, but we find that all affective states last longer in our data set than in their previous work.

The application of sensor-free detectors to the study of student affect provides the opportunity to study how such affect is exhibited in students at greater scale and at second-by-second levels of granularity. In addition, automated detectors are a less intrusive method of data collection than more traditional methods. As the detectors utilize only data recorded from computer-based systems, they can estimate a student's affective state without interrupting their work, as can be the case with self-reporting methods, and does not hold a risk of observer effects where students change their behavior due to the presence of a human coder. The method also does not require the use of additional technology such as physical and physiological sensors that may be difficult to deploy in classrooms at scale. Given the greater scale facilitated by automated affect detectors, future research may be able to study not just overall affective dynamics and chronometry but how dynamics and chronometry vary between different activities, different student populations, and even at different times of day. The better understanding of affective dynamics and chronometry that this may afford may have several benefits. Understanding a system's affective dynamics may be useful for encouraging positive transitions and suppressing negative transitions. Understanding affective chronometry may help us understand when negative emotion is problematic. Although some confusion is associated with positive learning outcomes [17], extended confusion is associated with worse student performance [18]. Understanding whether a student's confusion or frustration lasts longer than the expected duration may indicate that a student is struggling and is in need of intervention.

As the scale of the application of automated detectors increases for the study of affective dynamics, the means of evaluating common transitions will likely need to evolve as well. After a certain data set size, all transitions will become significant. Even in this paper, with a relatively limited data

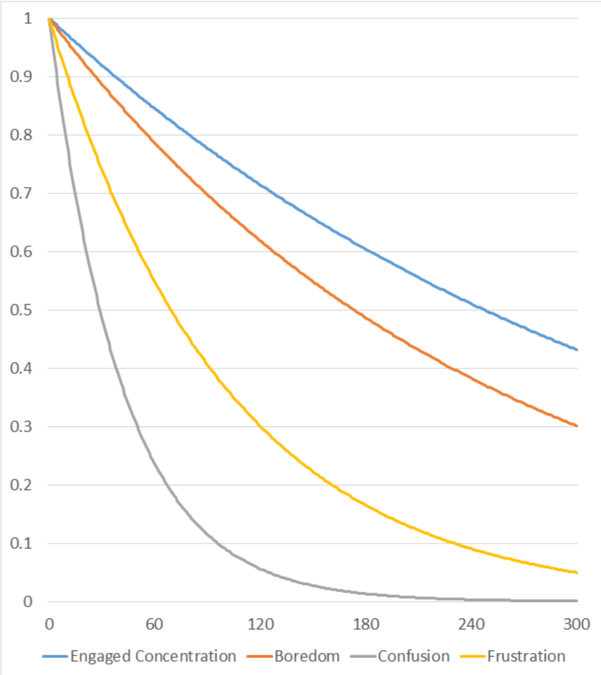

Figure 4: The plotted exponential decay of each affective state as estimated by the sensor-free affect detectors.

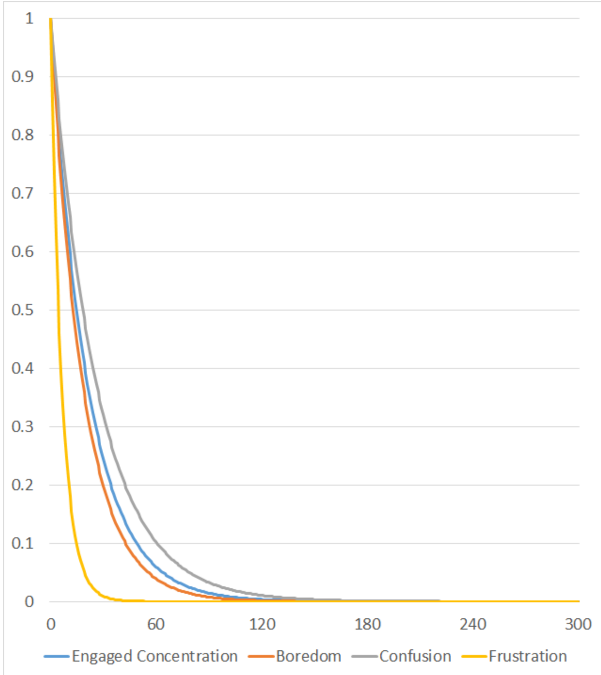

Figure 5: The plotted exponential decay of each affective state as reported in Table 1 of D'Mello and Graesser [9]

set, fairly low values of D'Mello's L reached statistical significance. Future work may need to explore new methods of identifying and evaluating affect dynamics, perhaps by simply exploring reasonable means of leveraging D'Mello's L as a measure of magnitude to identify meaningfully frequent links, not just those that are simply statistically significantly more likely than chance.

There are potential limitations to the current work that may be addressed by future research in this area. First, while the sensor-free detectors used in this work, as presented in [5], exhibit significantly superior performance to previous developed detectors with regard to AUC, improving the performance of these models further may help to improve transition and chronometry estimates, particularly of the less common labels of confusion and frustration. Utilizing methods to supplement less-frequently occurring labels of student affect (though the common method of resampling did not, in fact, enhance these detectors) or utilizing unlabeled data to better inform model estimates through co-training may improve model performance and produce more accurate measurements of affect dynamics and affective chronometry. It also may make sense to use different confidence thresholds for different affective states to adjust for the differences in the conservatism of different detectors that emerge from having different base rates.

Although consisting of a small portion of the data used in this work, the analyses did not include cases of co-occurring labels as estimated by the model. The estimates produced by the sensor-free detectors, even when the ground truth labels used to train such detectors did not observe co-occuring affective states themselves, is able to produce such cases, providing the opportunity to observe such cases in future work. Identifying which states are likely to co-occur, as well as include such cases in analyses of state transitions and affect state decay, will help to gain a better understanding of the relationships between affective states as well as to student performance.

A final opportunity for future work is in regard to observing affect dynamics and chronometry in experimental settings, as in the case of randomized controlled trials (RCTs). Several works have used analyses of state transitions to observe differences in affect exhibited between experimental conditions [27][8]. As the training set used to develop affect detectors does not contain experiment data, it is at this time uncertain if they generalize to behaviors exhibited outside of normal usage of the learning platform. Future work can observe how well such detectors generalize to such populations of users and samples.

6. ACKNOWLEDGMENTS

We thank multiple current NSF grants (IIS-1636782, ACI-1440753, DRL-1252297, DRL-1109483, DRL-1316736, DGE-1535428 & DRL-1031398) , the US Dept. of Ed (IES R305A120125 & R305C100024 and GAANN), and the ONR.

7. REFERENCES

[1] R. S. Baker, M. M. T. Rodrigo, and U. E. Xolocotzin. The dynamics of affective transitions in simulation problem-solving environments. In *International Conference on Affective Computing and Intelligent Interaction*, pages 666–677. Springer, 2007.

[2] Y. Benjamini and Y. Hochberg. Controlling the false

discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[3] N. Bosch and S. D'Mello. Sequential patterns of affective states of novice programmers. In *The First Workshop on AI-supported Education for Computer Science (AIEDCS 2013)*, pages 1–10, 2013.

[4] N. Bosch and S. D'Mello. The affective experience of novice computer programmers. *International Journal of Artificial Intelligence in Education*, 27(1):181–206, 2017.

[5] A. F. Botelho, R. S. Baker, and N. T. Heffernan. Improving sensor-free affect detection using deep learning. In *International Conference on Artificial Intelligence in Education*, pages 40–51. Springer, 2017.

[6] S. Craig, A. Graesser, J. Sullins, and B. Gholson. Affect and learning: an exploratory look into the role of affect in learning with autotutor. *Journal of educational media*, 29(3):241–250, 2004.

[7] M. Csikszentmihalyi. Flow. the psychology of optimal experience. new york (harperperennial) 1990. 1990.

[8] S. D'Mello and A. Graesser. Modeling cognitive-affective dynamics with hidden markov models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 32, 2010.

[9] S. D'Mello and A. Graesser. The half-life of cognitive-affective states during complex learning. *Cognition & Emotion*, 25(7):1299–1308, 2011.

[10] S. D'Mello and A. Graesser. Dynamics of affective states during complex learning. *Learning and Instruction*, 22(2):145–157, 2012.

[11] S. D'Mello, R. Taylor, and A. Graesser. Monitoring affective trajectories during complex learning. In *29th Annual Meeting of the Cognitive Science Society*, pages 203–208. Springer, 2012.

[12] T. F. G. Guia, M. M. T. Rodrigo, M. Dagami, C. Marie, J. O. Sugay, F. J. P. Macam, and A. Mitrovic. An exploratory study of factors indicative of affective states of students using sql-tutor. *Research & Practice in Technology Enhanced Learning*, 8(3), 2013.

[13] T. F. G. Guia, J. O. Sugay, M. M. T. Rodrigo, F. J. P. Macam, M. M. C. Dagami, and A. Mitrovic. Transitions of affective states in an intelligent tutoring system. *Proceedings of the Philippine Computing Society*, pages 31–35, 2011.

[14] N. T. Heffernan and C. L. Heffernan. The assistments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education*, 24(4):470–497, 2014.

[15] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[16] B. Kort, R. Reilly, and R. W. Picard. An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion. In *Advanced Learning Technologies, 2001. Proceedings. IEEE International Conference on*, pages 43–46. IEEE, 2001.

[17] B. Lehman, S. D'Mello, and A. Graesser. Confusion and complex learning during interactions with computer learning environments. *The Internet and Higher Education*, 15(3):184–194, 2012.

[18] Z. Liu, V. Pataranutaporn, J. Ocumpaugh, and R. Baker. Sequences of frustration and confusion, and learning. In *Educational Data Mining 2013*. Citeseer, 2013.

[19] M. Miserandino. Children who do well in school: Individual differences in perceived competence and autonomy in above-average children. *Journal of educational psychology*, 88(2):203, 1996.

[20] J. Ocumpaugh, J. M. Andres, R. Baker, J. DeFalco, L. Paquette, J. Rowe, B. Mott, J. Lester, V. Georgoulas, K. Brawner, et al. Affect dynamics in military trainees using vmedic: From engaged concentration to boredom to confusion. In *International Conference on Artificial Intelligence in Education*, pages 238–249. Springer, 2017.

[21] J. Ocumpaugh, R. Baker, S. Gowda, N. Heffernan, and C. Heffernan. Population validity for educational data mining models: A case study in affect detection. *British Journal of Educational Technology*, 45(3):487–501, 2014.

[22] Z. A. Pardos, R. S. Baker, M. San Pedro, S. M. Gowda, and S. M. Gowda. 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(1):107–128, 2014.

[23] B. C. Patrick, E. A. Skinner, and J. P. Connell. What motivates children's behavior and emotion? joint effects of perceived control and autonomy in the academic domain. *Journal of Personality and social Psychology*, 65(4):781, 1993.

[24] M. O. Pedro, R. Baker, A. Bowers, and N. Heffernan. Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. In *Educational Data Mining 2013*, 2013.

[25] G. N. Redondo. Turning points en las trayectorias emocionales de estudiantes en un contexto desafiante de aprendizaje experiencial: una aproximación dinámica. In *Quintas Jornadas de Jóvenes Investigadores de la Universidad de Alcalá: Humanidades y Ciencias Sociales*, pages 245–254. Servicio de Publicaciones, 2016.

[26] M. M. T. Rodrigo, R. d. Baker, J. Agapito, J. Nabos, M. Repalam, S. Reyes Jr, and M. San Pedro. The effects of an embodied conversational agent on student affective dynamics while using an intelligent tutoring system. *IEEE Transactions on Affective Computing*, 2(4):18–37, 2011.

[27] M. M. T. Rodrigo, R. S. Baker, J. Agapito, J. Nabos, M. C. Repalam, S. S. Reyes, and M. O. C. San Pedro. The effects of an interactive software agent on student affective dynamics while using; an intelligent tutoring system. *IEEE Transactions on Affective Computing*, 3(2):224–236, 2012.

[28] M. M. T. Rodrigo, R. S. Baker, M. C. Jadud, A. C. M. Amarra, T. Dy, M. B. V. Espejo-Lahoz, S. A. L. Lim, S. A. Pascua, J. O. Sugay, and E. S. Tabanao. Affective and behavioral predictors of novice programmer achievement. In *ACM SIGCSE Bulletin*, volume 41, pages 156–160. ACM, 2009.

[29] M. M. T. Rodrigo, G. Rebolledo-Mendez, R. Baker,

B. du Boulay, J. Sugay, S. Lim, M. Espejo-Lahoz, and R. Luckin. The effects of motivational modeling on affect in an intelligent tutoring system. In *Proceedings of International Conference on Computers in Education*, volume 57, page 64, 2008.

[30] M. O. San Pedro, J. Ocumpaugh, R. S. Baker, and N. T. Heffernan. Predicting stem and non-stem college major enrollment from middle school interaction with mathematics educational software. In *EDM*, pages 276–279, 2014.

[31] R. L. Solomon and J. D. Corbit. An opponent-process theory of motivation: I. temporal dynamics of affect. *Psychological review*, 81(2):119, 1974.

[32] Y. Wang, N. T. Heffernan, and C. Heffernan. Towards better affect detectors: effect of missing skills, class features and common wrong answers. In *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge*, pages 31–35. ACM, 2015.