Eli Pleaner, Julien Putz
Final project for EDUC C260
May 13, 2022

https://github.com/epleaner/mled-rqa

Multimodal Learning Analytics using Machine Learning and Recurrence Quantification Analysis

Introduction

Multimodal learning environments afford an abundance of observable interaction for informing learning science research. Among this interaction are haptic feedback and gaze, which, when analyzed via theoretical approaches such as embodied cognition, enactivism, and ecological dynamics (Abrahamson & Sánchez-García, 2016; Hutto et al., 2015; Varela et al., 1991), hold the potential to reveal important insights into the development of sensorimotor-facilitated conceptual learning. By analyzing how students' sensorimotor behavior shifts over time in response to task-based perception, behavioral patterns are revealed that indicate phase shifts in conceptual understanding, reflecting the complex and dynamic nature of embodied content-based learning (Abdu et al., under review; Tancredi et al., 2021; Tancredi et al., in press). In applying predictive machine learning techniques to telemetry data, multimodal learning analytics aims to contribute to research-informed education design, as well as improve educational support interventions such as adaptive tutoring systems.

The motivation for our research is to expand on previous approaches to gaze- and bimodal-based multimodal learning analytics in the context of an embodied design framework for learning proportionality. We seek to either reproduce or challenge expert-defined behavioral stages via unsupervised clustering using skip-grams and recurrent quantification analysis (RQA). Furthermore, we apply a long-short term memory recurrent neural network (LSTM RNN) in order to predict future behavior given a behavioral sequence, and, in a similar fashion, apply deep knowledge tracing (DKT) to predict future performance. Finally, we model this motor-coordination task as a skill-based situation, and explore the application of bayesian knowledge tracing (BKT) in predicting skill-based development over time. For each analytical technique, we report on our findings and discuss implications and limitations of our findings in educational interventions.

Dataset

This research works with data collected from the Mathematics Imagery Training for Proportionality (MIT-P), an action-based embodied design framework (Abrahamson, 2014) for

grounding proportionality understanding in sensorimotor schema developed through task-based motor-control coordination. On the screen are two vertical bars, extending from the base of the screen. The height of each bar is determined by the last touched location for the left and right hand. The instruction participants receive is to “make the bars green”. The bars are green when the height of the two bars are at a 1:2 ratio, and are red otherwise. Participants are not given this information, and thus must explore the task space to discover this rule of proportionality.

This data was gathered from a previous research study with students from the Netherlands in fifth and sixth grade. This data consists of 14 student trials, including eye-gaze and bimanual touchscreen coordinates, downsampled to 10Hz. These trials ranged in length from 3:16 minutes to 5:24 minutes. Each row consisted of telemetry data for one student at a 0.1 second time slice.

	TIME[s]	LeftX	LeftY	RightX	RightY	GazeX	GazeY	GazeDuration	id
0	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	01
1	0.1	NaN	NaN	NaN	NaN	374.0	285.0	1451.0	01
2	0.2	NaN	NaN	NaN	NaN	374.0	285.0	1451.0	01
3	0.3	NaN	NaN	NaN	NaN	208.0	145.0	369.0	01
4	0.4	NaN	NaN	NaN	NaN	208.0	145.0	369.0	01
...	...	...	...	...	...	...	...	...	...
41671	354.4	118.0	482.0	613.0	958.0	591.0	976.0	1717.0	33
41672	354.5	123.0	488.0	613.0	962.0	591.0	976.0	1717.0	33
41673	354.6	125.0	493.0	612.0	967.0	591.0	976.0	1717.0	33
41674	354.7	125.0	495.0	612.0	971.0	591.0	976.0	1717.0	33
41675	354.8	127.0	498.0	612.0	977.0	591.0	976.0	1717.0	33

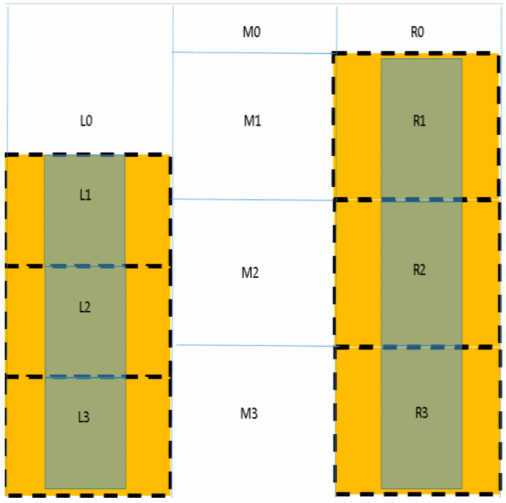
41676 rows × 9 columns

Figure 1. Sample rows from dataset

Related Work

Our work expands on that of Tancredi et al. (2021; Pardos et al., 2022), which introduces a proof of concept for applying RQA to bimanual telemetry data for modeling phases of understanding within the MIT-P context. Using expert-labeled stages of behavior (Exploration, Discovery, and Fluency), this paper demonstrates that there are signals within the RQA metric for predictability that correspond to these stages of behavior, as well as the transition between them. By analyzing a RQA plot, one can discern distinct features of behavioral recurrence that align with the intuitions behind the expert-labeled stages of behavior, suggesting that RQA may be used to characterize and predict critical developmental shifts in learning (Tancredi et al., 2021).

We also draw heavily on Adbu et al.'s (in press) research which used the same dataset as the present study. Their research models multimodal mathematical learning as a complex dynamic system in a similar manner as Tancredi et al. (2021). They again document stages of behavior using RQA metrics. However, instead of only leveraging one modality (bimanual telemetry data), Adbu et al. focus on eyes-hands dynamics, combining bimanual data with gaze tracking data. In particular, they combine both measures into one feature, which they term "Areas of Interest" (AoIs, see Figure 2). The intuition for using this feature is that the absolute gaze position at the top of the screen can mean something very different depending on whether the hands are close to the top of the screen as well, or close to the bottom. The lateral boundaries of the three bars (Figure 2) are defined for each student individually, relative to the median position of each hand. The subdivision of each column into AoIs changes dynamically at each time slide, relative to the height of each hand position. Using these dynamic AoIs serve as input for doing RQA, the authors document significant changes in RQA metrics across stages of behavior—a finding that is consistent with the work of Tancredi et al.



Note. The Trainer screen, allocated to AoI. L1-L3 and R1-R3 represent AoI on the left and right bars (and their surroundings). L0 and R0 are the areas above the bars, and M0-M3 represent the area between the bars.

Figure 2. Coding Scheme for Areas of Interest

Next, we follow Pardos et al. (2022) who apply Recurrent Neural Networks (RNNs) to bimanual telemetry data collected in the context of MIT-P. In particular, the authors used Long Short Term Memory networks (LSTMs) to predict student strategies, and they improved upon the performance of a baseline that used logistic regression. Further, the authors use a simple RNN model as part of a visualization tool, revealing behavioral patterns detected by the RNN.

Furthermore, our work draws on the work of Huang et al. (Huang et al., 2019) who combine gaze-tracking, electrodermal activity, and kinesthetics for a multimodal learning analytic approach to classifying collaborative learning states. This work applied unsupervised clustering

techniques to discover distinct behavioral states without the need of manual, expert-level labeling. This research also analyzed the transition probabilities across these states as a function of time.

Finally, we draw on theoretical constructs from Abrahamson and Sánchez-García (2016), most notably their notion of attention anchors. This enactivist notion seeks to explain how learners develop motor-action skills in motor-control problems. More specifically, an attentional anchor is "a real or imagined object, area, or other aspect or behavior of the perceptual manifold that emerges to facilitate motor-action coordination" (p. 203). In the context of MIT-P, an example of an attentional anchor is the vertical interval between the left and the right hand. In order to fluently move "in green", a student can visually focus on this interval in their perceptual field and employ the following heuristic: "The interval must become bigger as my hands become higher". By following this heuristic, the student no longer needs to attend to two perceptual structures (i.e. the two hands) but only to a single one (i.e. the interval), thus favorably collapsing two motor-action schemes into a single scheme. Another example of an attentional anchor is a perceptual triangle, which is illustrated in Figure 3. Research learning behavior in the context of MIT-P has documented multiple such attentional anchors, which usually correspond to visually focussing on actual or imagined objects in the perceptual field.

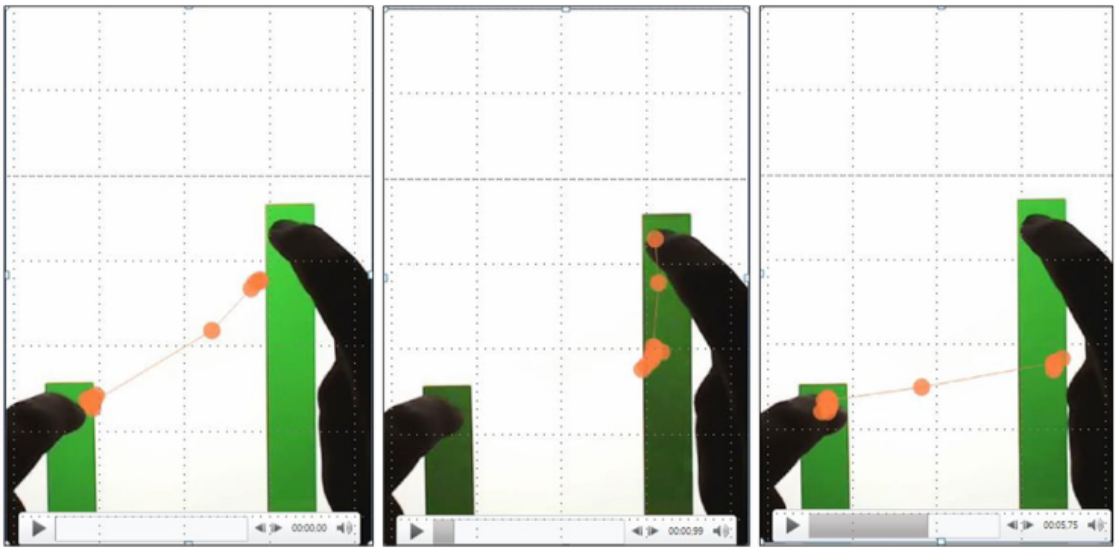


Figure 3. Illustration of a perceptual triangle captured through gaze data (Duijzer et al., 2017). The student looks at the top of the short bar, top of the high bar and halfway up the high bar.

Methods and Results

Areas of Interest

We made use of the Areas of Interest presented by Abdu et al. (in press). We recreated these areas using Python code. For each student, we first defined the left (resp. right) border of the screen by computing the median and range of the left (resp. right) hand x-coordinates and by subtracting (resp. adding) the range from the median. The screen was then partitioned into three vertical bars of equal width. Horizontal allocation was done for each column separately. At each 0.1sec timestamp, we split the left (resp. right) bar below the current left (resp. right) hand position into three equal areas L1, L2, L3 (resp. R1, R2, R3). Following Abdu et al., we multiplied the vertical hand coordinates by 1.1 to account for spatial variance in eye gaze. The area above the left (resp. right) bar was labeled L0 (resp. R0). The middle column was split into M0, M1, M2, M3 according to the bar that was currently higher. For instance, if the right hand was higher than the left hand at a particular timestamp, then the middle column would be split into four parts according to the right column (see Figure 2).

Skip-gram visualizations

In the exploratory stage of our study, we evaluated the utility of word2vec models for visualizing patterns in our dataset. In particular we applied skip-gram, an unsupervised learning technique used to predict the context of a word in text. We prepared the data by rounding gaze position and hand height to the nearest 50 (e.g. 441.3 becomes 450.0) and by concatenating the resulting numbers into a string. The format of these strings was "LeftY RightY GazeX GazeY", where LeftY and RightY are the heights of the left and right hands, respectively, and where GazeX and GazeY represent the x and y coordinates of the student's gaze. For each student, we put these strings into a chronological list: ["LeftY_1 RightY_1 GazeX_1 GazeY_1", "LeftY_2 RightY_2 GazeX_2 GazeY_2", "LeftY_3 RightY_3 GazeX_3 GazeY_3"...]. In this way, each word represents approximate hand-eye positions at a particular time point, and the context of each word corresponds to preceding and subsequent hand-eye positions. The reason why we rounded the coordinate numbers to the nearest 50 was to obtain a reasonable ratio between the total number of words (41,676) and the number of distinct vocabulary (9,064). We then trained a skip-gram model on these lists and visualized the hidden layer using t-SNE. We first color-coded the t-SNE visualizations according to whether the corresponding hand positions are "in green"; that is, the right hand is approximately twice as high above the bottom of the screen as the left hand. The results are shown in Figure 3, left. We then color-coded the embedding according to the Areas of Interest (AoIs) outlined above; that is, colors indicate that the gaze position is located within a particular AoI. The results are shown in Figure 3, right.

For the “in green”-coded visualization, we observe a clear separation between words corresponding to “being in green” (shown in blue in Figure 3, left) and words corresponding to not “being in green” (shown in orange). This suggests that the skip-gram model can pick up on patterns in the data, even when the data is treated as a string rather than as a set of numbers. For the AoI-coded visualization, we do not observe any overt patterns beyond slight overrepresentation of L0 points in the top right and a slight overrepresentation of R0 points in the bottom left (see Figure 3, right). This suggests either that skip-gram is not a sensitive enough method for picking up on the AoIs in the data, or that AoIs are not the most meaningful or natural way of partitioning our data. Further research is needed to fully explore the affordances of word2vec-based visualization techniques, and to make comparisons with other techniques such the those presented by Pardos et al. (2022).

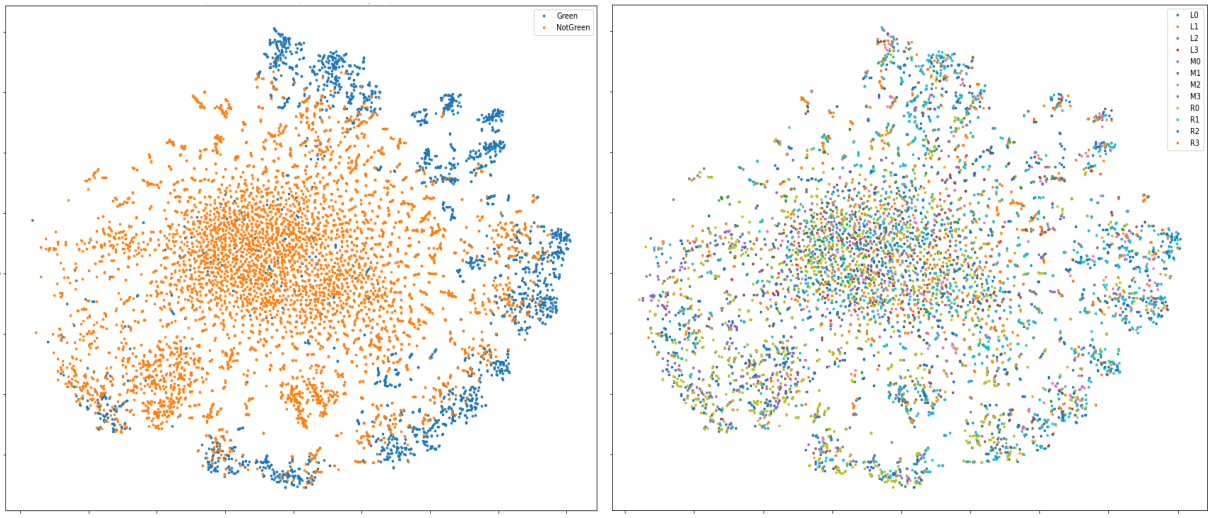


Figure 4: t-SNE visualizations of hand-graze position skip-gram embedding. Left figure is coded according to “being in green”. Right figure is coded according to Areas of Interest.

RQA K-Means clustering

In order to apply unsupervised clustering to our time-series data, we used RQA metrics as features. To begin, we forward-filled all null values for bimodal and gaze coordinates, which consisted of around 10% of the dataset. We used this approach, rather than dropping these rows, in order to avoid a discrepancy around the number of time-slices in each student time-series. For each student, we split the time-series into many short time windows, and calculated the RQA metrics for each window using pyrqa[1]. We used pyrqa’s cross-RQA analysis method in order to calculate recurrence rate, determinism, average diagonal line length, entropy, and trapping time. We then created a dataframe with each row being these RQA metrics for a particular time window for a particular student. We clustered this data using scikit-learn’s KMeans, plotting the silhouette score for k=[2, 20] clusters and applying the elbow method to determine the

[1] https://pypi.org/project/PyRQA/

best-fitting value for k. We then used scikit-learn's TSNE to reduce the dimensionality of this RQA metric data from five dimensions to two, in order to visualize the clustering via d3-scatterplot[2]. Finally we repeated the same clustering approach but with RQA metrics calculated for each student across their entire task time-series.

Figure 5 shows visualizations of clusters based on RQA metrics aggregated at different time windows.

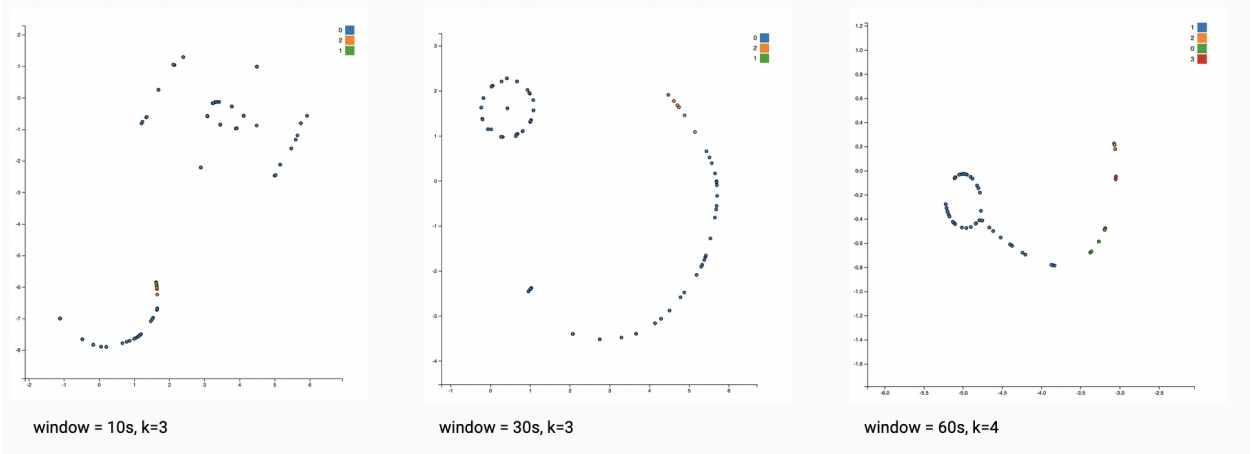


Figure 5. Cluster results based on RQA metrics for different time windows

As these clustering results do not indicate a clear signal of distinct patterns of RQA metrics, they convey the difficulty in clustering based on these metrics. When applying this windowing approach, there emerged tradeoffs between the size of the window (resulting in more meaningful RQA metrics) and the amount of resultant data points to cluster with. Furthermore, across time windows, the RQA metrics were heavily left-skewed towards zero. From this there emerged one predominant cluster, and the few outliers were split into two or three other clusters.

Figure 6 shows the visualization of clustering (k=5) based on RQA metrics across each student's trial. When clustering across entire trials, clustering failed to produce meaningful results. As there were 14 students, each with distinct RQA metrics for their trial, clustering resulted in no signal with regards to data grouping.

[2] https://github.com/CAHLR/d3-scatterplot

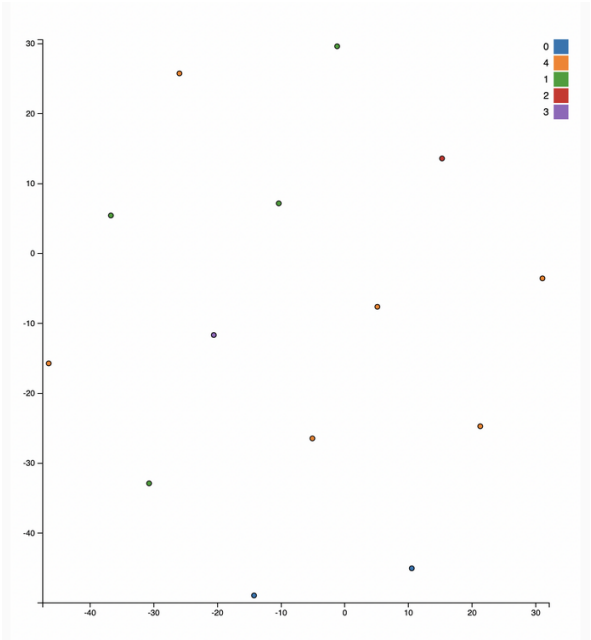


Figure 6. Cluster results based on RQA metrics for different time windows

We were surprised to see so little signal emerge when clustering based on RQA metrics. We believe that this approach was limited by the available data, including both a short duration of trials as well as the small sample of students. Our hypothesis was that RQA metrics might be an indicator of learning development, as well as of distinct learning styles within the student population. In this way, we hoped to validate or expand on previous work showing RQA patterns within expert-labeled stages of learning. We believe that with a dataset with more students and longer trials, applying this same RQA-based clustering approach may yield more significant results.

Behavior prediction using LSTM RNN

We began by taking only the first n time-slices for all students, where n is the length of the shortest student trial. This reduced our dataset from 41676 to 27425 rows, a reduction of 34.19%. We then calculated the average difference between the target bimanual proportionality and the actual for every second for each student across their time-series. This was done by finding the ratio of the right hand y-coordinate to the left, and subtracting 2 (representing a 2:1 target ratio). We then binned these windowed differences into eight bins, corresponding to the distribution of the differences across the dataset. Our binning followed this rule:

Average difference of proportion (1 second window)	Bin	Count
< 0.1	0	185
< 0.3	1	307
< 0.5	2	179
< 0.7	3	809
< 1	4	432
< 2	5	758
< 3	6	17
>= 3	7	43

Table 1. Bin labels for windowed average difference of actual proportion to target

We then modified the shape of this resultant dataset such that each column corresponded to the bin label for one time window (e.g 1 second – 2 second), and each row corresponded to one student. We then split this dataset into train and test sets, with a 65%–35% respective split. We then built a LSTM RNN model using Keras, determining the optimal hyperparameters via a grid search to be *lstm_dim = 128, batch_size = 16,* and *validation_split = 0.2*. Our evaluation metric was root mean squared error.

We used as a baseline metric a predictor that would always predict the average distribution of bins for any given time window, which gave an RMSE=0.347. Our RNN performed with RMSE=0.304 – an improvement of 12% over the baseline. Our results aligned with our expectations for this given task, and indicate that a LSTM RNN enables accurate behavioral predictions when trained on one modality. Still to be done is an exploration of the incorporation of gaze tracking into this behavioral feature space.

As our task tightly couples performance and behavior (target proportionality is a direct result of hand positions), we distinguish this behavioral prediction model from that of performance predictions by using this model to predict more fine-grained levels of behavioral performance. This model enables us not only to predict in-green or not-in-green (as the subsequent DKT model does), but to predict categorical levels of correctness: how close a participant will be to the target proportion, given past behavior. While we make a distinction between behavioral prediction (with more fine-grained categorical levels) and performance prediction (with a binary correctness measure), we acknowledge that the distinction between performance and behavior is somewhat fluid in the present context.

Performance prediction using LSTM RNN

Previous research has shown promising results for using RNNs in one modality (e.g. bimanual telemetry data). We wanted to build on this work by leveraging data taken across multiple modalities. We used the AoIs outlined previously as a feature combining gaze tracking with bimanual telemetry data. Our goal was to predict "being in green" from AoIs. As before, we only used the first n time-slices for each student (n being the length of the shortest student trial), and we split this reduced dataset into train (~65%) and test sets (~35%). We then built a LSTM model using Keras.

The model had the same architecture as Deep Knowledge Tracing (DKT; Piech et al., 2015); however, instead of skills we used AoIs, and instead of response correctness we used "being in green". The input is a one-hot encoding of the students' interaction that represents the combination of which AoI the student looked at and if the student was "in green". The output is a binary prediction probability of being "in green" in the subsequent time-slice. After doing a hyperparameter sweep, we set *lstm_dim = 20*, *dropout = 0.2*, and *epochs = 0.2*. Our evaluation results are summarized in Table 2. In terms of prediction accuracy and RMSE, our model outperformed a simple baseline which at each time-slice predicted the percent of "in green" instances in the training dataset (41.60%).

	LSTM (DKT architecture)	**Baseline** (using percent of "in green" instances)
Accuracy	81.86%	58.40%
RMSE	0.3646	0.4948

Table 2: Evaluation results for the LSTM model based on the DKT architecture

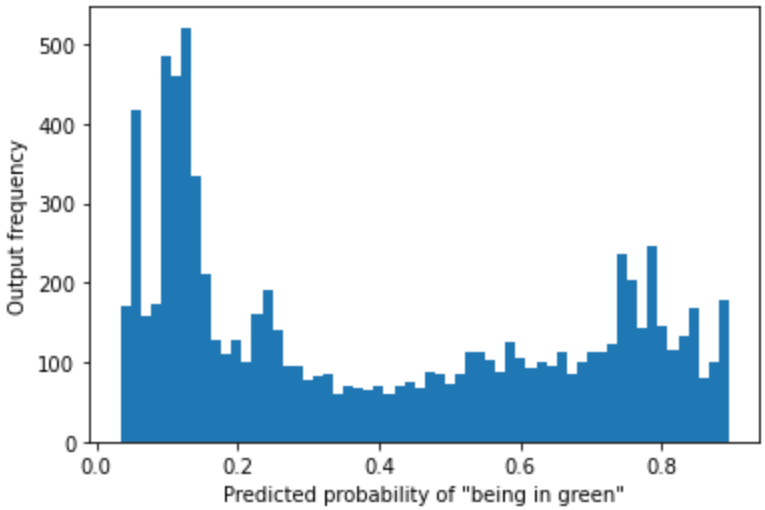


Figure 7. Distribution of LSTM/DKT output probabilities across all students and timesclices in the test dataset.

Skill modeling using BKT

Neural networks such as RNNs afford strong prediction performance and generally do not require explicit encoding of human domain knowledge. Correspondingly, we applied different types of RNNs to our dataset without committing to strong theoretical assumptions about knowledge and learning. A downside of this approach is a lack of interpretability; the weights in the above RNN models are difficult to scrutinize and to relate to learning theory. An alternative approach to modeling learning that has long been used in the context of interactive tutoring systems is Bayesian Knowledge Tracing, or BKT (Corbett & Anderson, 1994; Yudelson et al., 2013). BKT uses four hidden parameters—prior knowledge, learning rate, guess, and slip—which have an intuitive and theoretically grounded meaning. While our dataset is qualitatively different from the kinds of data that BKT is traditionally used for, we draw on the work of Abrahamson and Sánchez-García (2016) to justify the application of BKT to our multimodal data.

First, we propose that looking at a particular AoI while moving “in green” can be interpreted as leveraging a particular attentional anchor in order to solve the motor control task of MIT-P. In order words, we operationalize attentional anchors through AoIs. Second, we propose that each distinct attentional anchor (and hence each AoI) corresponds to a distinct motor-action skill; if a student manages to move “in green” while visually attending to the vertical interval between the hands, this is a different skill than moving “in green” while attending to the perceptual triangle (see Figure 3). Third, mastering a particular skill corresponds to successfully moving “in green” while leveraging a particular attentional anchor. In short, we view performance in the MIT-P environment in terms of distinct skills, each with a binary mastery parameter. These are precisely the assumptions underlying BKT.

In light of the above discussion, we believe we are theoretically justified to apply BKT on our dataset. We use pyBKT (Badrinath et al., 2021) to train a BKT model, using the same 65% training data set as before. We evaluate our model on the 35% test dataset, the results of which can be seen in Table 3. Again, we greatly outperform the percent-based baseline. However, the prediction of our BKT model is slightly inferior to our DKT/LSTM model. That being said, given the higher interpretability and lower computational cost, this may be regarded as a worthwhile trade-off—at least in some contexts. In interactive tutoring systems, for instance, BKT models may be more appropriate as domain experts can leverage the interpretable parameters to inform curriculum development, while the computation efficiency makes real-time performance modeling feasible.

	BKT	**Baseline** (using percent of "in green" instances)
Accuracy	80.72%	58.40%
RMSE	0.3796	0.4948

Table 3: Evaluation results for the BKT model

Next, we inspected the parameters of the trained BKT model, focusing primarily on learning rate. Figure 8 shows the learning rates for each AoI. The areas with the highest learning rate, R2, corresponds to halfway up the high bar and is one of the key areas associated with the attentional anchor that we introduced earlier as the "perceptual triangle". This suggests that this particular attentional anchor may have played an important learning role in the present dataset. The second highest learning rate is associated with area M2. This area is arguably part of the vertical interval between the two hand positions, which previous research has identified as another important attentional anchor for solving the MIT-P task. This finding is consistent with previous qualitative research that has identified these two attentional anchors as important factors for learning within the MIT-P environment (Abrahamson & Sánchez-García, 2016; Duijzer et al., 2017).

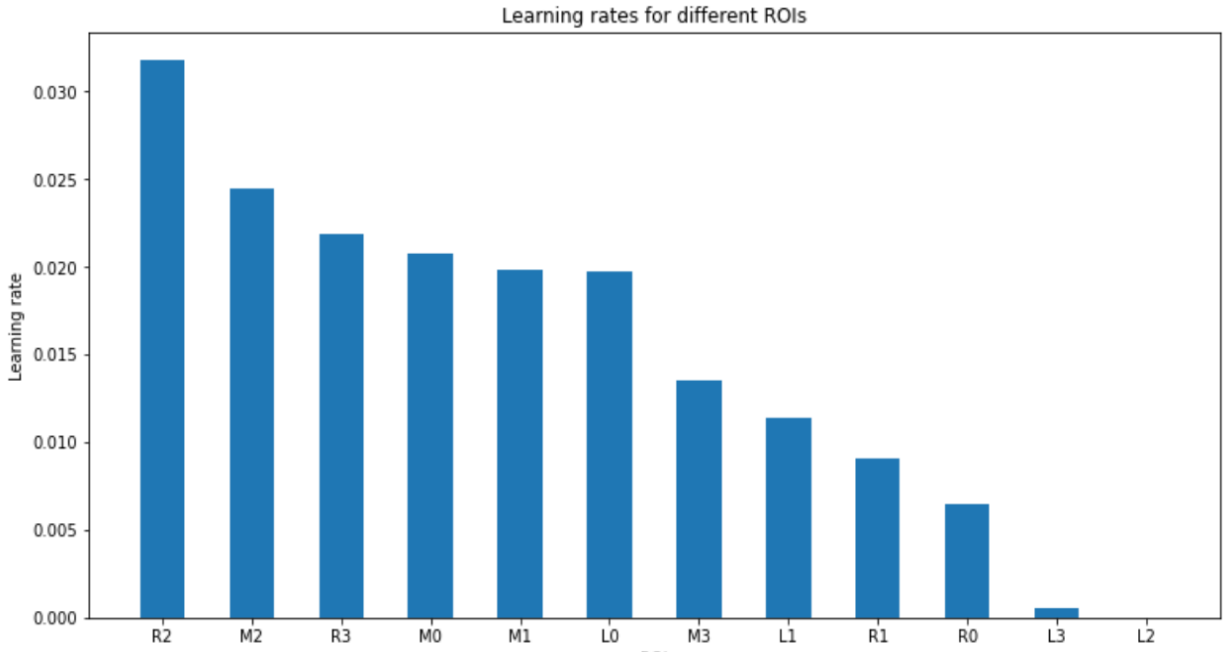


Figure 8. BKT learning rates for each AoI

Conclusion and Future Research

In this study, we applied various machine-learning and dynamic systems techniques to multimodal data collected from the MIT-P. Our work opens up a number of directions for future research. First, the dataset we used was relatively small, with a relatively low temporal resolution by gaze tracking and telemetry standards. We hope to apply the present methods to a larger dataset with higher temporal resolution. In particular, we believe that with a larger dataset, our RQA-based clustering approach may yield more significant results. There is potential in this larger dataset to also incorporate further modalities, including pupil dilation or electroencephalographic (EEG) as proxies for attention and affect, as the analytical techniques used here are not exclusive to any particular modality – the crucial task here would be to produce multimodal features analogous to the haptic-gaze AoI.

Second, our use of AoIs as proxy for attentional anchors requires further validation and refinement. It is unlikely that the present AoIs optimally capture attentional anchors. For instance, we interpreted area R2 as a proxy for the "perceptual triangle", even though R1 and L1 may also plausibly be regarded as part of this particular attention anchor. Similarly, we used area M2 as a proxy for the vertical interval between both hands, even though M1 may also plausibly be part of this attention anchor. Future research could investigate whether theory-driven refinement of AoIs could lead to improved prediction accuracy in our models.

Third, we are curious about the possibility of using RQA-based features instead of windowed hand-height differences or AoIs in our RNN models. Finally, the methods presented in this study

may help augment MIT-P with further interactive tutoring capabilities. For instance, our performance prediction models (LSTM/DKT and BKT) may be used to generate real-time hints based on particular levels of mastery and particular AoIs. In all, this paper demonstrates promising efficacy for combining multimodal learning analytics with dynamic systems techniques for offline learning science research and for real-time adaptive tutoring systems in motor-control tasks.

References

Abdu, R., Tancredi, S., Abrahamson, D., & Balasubramaniam, R. (under review). A complex-systems view on mathematical learning as hand–eye coordination. *Educational Studies in Mathematics,* (Eye-tracking research in mathematics education [Special issue])

Abrahamson, D. (2014). Building educational activities for understanding: An elaboration on the embodied-design framework and its epistemic grounds. *International Journal of Child-Computer Interaction, 2*(1), 1-16. https://doi.org/10.1016/j.ijcci.2014.07.002

Abrahamson, D., & Sánchez-García, R. (2016). Learning is moving in new ways: The ecological dynamics of mathematics education. *Journal of the Learning Sciences, 25*(2), 203-239.

Badrinath, A., Wang, F., & Pardos, Z. (2021). pyBKT: an accessible python library of Bayesian knowledge tracing models. *arXiv Preprint arXiv:2105.00385,*

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction, 4*(4), 253-278.

Duijzer, C. A., Shayan, S., Bakker, A., Van der Schaaf, Marieke F, & Abrahamson, D. (2017). Touchscreen tablets: Coordinating action and perception for mathematical cognition. *Frontiers in Psychology, 8*, 144.

Huang, K., Bryant, T., & Schneider, B. (2019). Identifying Collaborative Learning States Using Unsupervised Machine Learning on Eye-Tracking, Physiological and Motion Sensor Data. *International Educational Data Mining Society,*

Hutto, D. D., Kirchhoff, M. D., & Abrahamson, D. (2015). The enactive roots of STEM: Rethinking educational design in mathematics. *Educational Psychology Review, 27*(3), 371-389.

Pardos, Z. A., Rosenbaum, L. F., & Abrahamson, D. (2022). Characterizing learner behavior from touchscreen data. *International Journal of Child-Computer Interaction, 31*, 100357.

Tancredi, S., Abdu, R., Abrahamson, D., & Balasubramaniam, R. (2021). Modeling nonlinear dynamics of fluency development in an embodied-design mathematics learning environment with Recurrence Quantification Analysis. Paper presented at the (100297)https://doi.org/10.1016/j.ijcci.2021.100297

Tancredi, S., Abdu, R., Balasubramaniam, R., & Abrahamson, D. (in press). Intermodality in multimodal learning analytics for cognitive theory development: A case from embodied design for mathematics learning. In M. Giannakos, D. Spikol, D. D. Mitri, K. Sharma, X. Ochoa & R. Hammad (Eds.), *Multimodal learning analytics* ()

Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience*. MIT press.

Yudelson, M. V., Koedinger, K. R., & Gordon, G. J. (2013). Individualized bayesian knowledge tracing models. Paper presented at the *International Conference on Artificial Intelligence in Education,* 171-180.