Predicting Behaviors based on Sequence Modeling of Test-takers' Clickstreams using LSTM, RNN, and n-gram

Steven Tang, Zhen Li

eMetric LLC

Paper written for the 2023 meeting of the National Council on Measurement in Education. It is prepared as part of the coordinated session on: Cheating Detection Using Machine Learning and Deep Learning Methods. The views expressed in this paper are solely those of the authors and they do not necessarily reflect the positions of eMetric LLC.

Correspondence concerning this paper should be addressed to Steven Tang, eMetric, 211 N Loop 1604 E, Suite 170, TX 78232. Email: steven@emetric.net.

Abstract

In large-scale computer-based assessment, clickstreams capture the exact clicks and behaviors of each test-taker throughout the exam period. In this study, several approaches towards predicting behavior in a test environment are analyzed with the purpose of quantifying how typical (or atypical) a student's behaviors are in a test context, providing a summary measure of a test-taker's behaviors, allowing for further investigation of any test-takers who are displaying atypical behavior patterns. The proposed behavior models include architectures such as the Long Short-Term Memory (LSTM) network, Recurrent Neural Networks (RNN), and an n-gram approach. The proposed models will predict the next action in a clickstream sequence given prior history. Model results will be evaluated using Model Agreement Index (MAI), a summary statistic of quantifying model agreement. Lower MAI score indicates fewer typical test-taking behaviors. Clickstream data is obtained from a state-wide summative test administered to grades 3-8 students in 2021. The characteristics of MAI indexes, the comparison among different prediction models, and correlations between MAI results and other existing statistics for detecting aberrant test-taking behaviors are discussed.

Key Words: Predictive Behavior Modeling, Clickstream, Model Agreement Index

Introduction

In a perfect testing scenario, test-takers fully represent their capabilities and knowledge by answering each test item in a test, and the resulting scores are an accurate representation of the test-takers' abilities. In practice, a variety of potential issues can arise. For example, test-takers could voluntarily undermine the testing process through cheating or refusing to authentically try their best. Additionally, the actual delivery of test content and items can vary from environment to environment depending on software, and sometimes students could be confused in how to correctly navigate the test or how to use tools available to them, which could negatively affect the test-takers' performance.

In this study, we propose the use of "predictive behavior modeling" to summarize the behavior patterns of test-takers by their clickstream data as a method to identify potential issues arising during the testing process. With these behavior models, a **Model Agreement Index (MAI)** is established. Lower values of MAI indicate that the clickstream contains actions that are atypical and harder to predict. Once clickstreams with low MAI have been identified, qualitatively and quantitively analyzing "why" such clickstreams are hard to predict can help stakeholders verify whether these sources of possible aberrance are acceptable or not. The underlying reasons why clickstreams have low MAI could vary for different testing administrations, as test content and test-taker populations vary.

Three prediction models are analyzed in this study. The first model analyzed is the Long Short-Term Memory (LSTM) network, a popular deep learning model applied to sequence data. The LSTM approach is compared to two baseline models: a vanilla recurrent neural network (RNN) and a bigram model. The use of the LSTM historically achieved state-of-the-art results in language modeling tasks (Sundermeyer, Schlüter, & Ney, 2012), which involve predicting the next word given prior context. The concept behind "predicting the next word in a sequence" can be analogous to "predicting the next behavior or action in a test-taking sequence," which is part of the motivation behind using the LSTM for the purpose of predicting test-taker behaviors.

The goal of applying these models is to give a straightforward quantification (MAI) of how typical an examinee's behaviors are within a testing context. The sequence behavior models are trained on clickstream data that includes all trackable actions in a computer-based test environment, including navigations, multiple-choice response selections, tool usage like calculator or notepad, and accommodations such as screen contrast toggling. The goal of each model is to predict the next clickstream action given the history of prior actions.

Operational Definition of Atypical Behavior

Suppose that a predictive model of student test-taking behaviors exists, with inputs being past clickstream actions and outputs being possible future actions. With this predictive model, one can define an "atypical clickstream" to be a clickstream that is not well predicted by the proposed model by comparing each observed action in the clickstream to the predicted probability of that observed action by the model's output. Clickstreams that are better predicted by the model are supposedly more "typical" as they are more predictable. In this study, three predictive models of behaviors based on a bi-gram, simple RNN, and LSTM architecture are proposed. The predictive models are then used to compute a Model Agreement Index (MAI) value, which indicates the extent of agreement between observed clickstream actions and model-predicted actions on a likelihood continuum ranging between 0 and 1. Clickstreams with relatively low MAI values are operationally considered more atypical than clickstreams with higher MAI values.

An assumption inherent to this study is that such a predictive model can be generally useful to stakeholders interested in ensuring that typical test-taking operations are observed, and that this model could serve as a system to monitor behavior patterns at scale, focusing on the entirety of a test rather than individual item responses. Monitoring algorithms are intended to flag noteworthy results to some degree of accuracy. For testing, noteworthy events could include “cheating behaviors” and “confusion.” It can be challenging to design these monitoring algorithms, as descriptions and signals of the cheating phenomenon and of student confusion are not precisely defined and may be extremely rare in practice. The operational definition of atypical in this paper serves as one lens in identifying “typical” and “atypical” behaviors, with the goal that flagging atypical behaviors using this definition will ultimately add value to stakeholders who want to ensure that typical test-taking processes are observed, and that atypical behaviors can be further analyzed to ensure nothing unwanted is occurring.

Related Work

Clickstream analysis has historically been used to determine and summarize user behaviors in web usage contexts (Banerjee & Ghosh, 2011; Heer & Chi, 2002). In these works, users’ navigation paths within a website were analyzed to obtain information about users’ preferences. Clustering techniques have been used to group together clickstreams with similar behavior usage patterns (Gunduz & Ozsu, 2003; Su & Chen, 2015); these clusters were used to infer user interests and predict future user behaviors. LSTMs trained on clickstream data have been used to predict student navigational pathways (Tang, Peterson, & Pardos, 2017) in massively open online course environments. In terms of aberrant and malicious user detection, clickstream analysis has been used to detect potential attackers who create fake identities in social media platforms (Wang, et al., 2017). In that work, sub-sequence counting with clustering is used to categorize clickstreams into different user archetypes, identifying clusters of clickstreams that could potentially be flagged for banning in their respective social media platforms.

In the field of educational testing, clickstreams (A.K.A, process data) have attracted more attention in recent years coinciding with the rise in popularity of online testing. K-means clustering was applied to process data for extracting behavior patterns of test-takers when they are measured on problem-solving skills (He, Liao, & Jiao, 2019). In addition, two recent approaches were developed to extract latent features from action sequences (Tang, Wang, He, Liu, & Ying, 2020; Tang, Wang, Liu, & Ying, 2020). Two underlying models, multidimensional scaling (MDS) and sequence-to-sequence autoencoders, are used to capture the pairwise dissimilarity of action sequences in process data. These features were found to be useful in predicting the final response of the test-takers for problem-solving items. Moreover, quite a few existing data forensics methods utilize one specific aspect of clickstream data at one time, e.g., examining if an item-response pattern is congruent with a specified measurement model (Drasgrow, Levine, & Williams, 1985), identifying extremely short or aberrant response times (Li, Wall, & Tang, 2018; van der Linden & Guo, 2008; Wise & DeMars, 2006), or detecting a large number of wrong-to-right answer changes at a group or individual level (Bishop & Egan, 2017). Recently, a new approach utilized multiple features like response times, number of actions, number of answer changes to identify the examinees whose test-taking processes deviate from most examinees (Liao, Patton, Yan, & Jiao, 2021). They discovered several archetypes of test-taking processes by applying k-means clustering algorithm. For example, an archetype can be a type of behavior that, comparatively, has long mean response time, many answer changes, and moderate variation in response time.

Dataset

The dataset for this study consists of clickstream data from a state-wide summative test administered to grade 8 students in 2021. Each row in the clickstream log contains key pieces of information: timestamp, click_action, user_id. The click_action is the actual click or action that was taken. The user_id identifies which test-taker produced the clickstream.

Table 13 in the appendix shows the 151 possible actions from this clickstream dataset. The approach in the current study has a larger, more complex input space compared to other approaches. The key benefit of using this more complex input space is that every instance of clickstream behavior is modelled, allowing the LSTM model to potentially learn many different patterns of test-taking behaviors.

Dataset Sample

The dataset used in this study consists of 3,934 Grade 8 examinee records, with a total of 531,628 clickstream rows, from the administration of a state-wide summative assessment in 2021. The 3,934 records represent every “valid” clickstream that was able to be processed for all students in one test session on one test form.

Methodology

For this study, each of the three predictive models is given as input the history, in sequential order, of behaviors that have occurred up to the current time point. The model is tasked with outputting a probability distribution for the action that could come next given this input history. The simple RNN and LSTM approaches are given the entire history of actions so far, while the MCNA model is effectively given a history of just the preceding action. This section provides a description of each of the three predictive approaches: RNN, LSTM, and MCNA.

Simple RNN

Recurrent neural networks (RNN; Graves, 2014) are neural networks with loops in them, allowing information to persist. The output from the previous step becomes the input to the next step, allowing for historical context to influence future predictions. This model is commonly applied to sequential data, such as language modeling or time series analysis. A simple RNN model consists of an input layer, a hidden layer, and an output layer.

Table 1 Hyperparameters for Simple RNN

Factors	Levels
batch_size	8, 32
epoch	0-99
lstm_node_size	128
layers	1
dropout	0.01
optimizer	‘Adam’

For this study, the RNN was implemented in Keras (Chollet & Others, 2015), an open-source software library that provides a Python interface for artificial neural networks with the machine learning library TensorFlow (Abadi, et al., 2015) serving as the back end. RNN models have a variety of hyperparameters that can be tuned. In the current study, most of the hyperparameters were selected based on the authors' experience in previous research (Tang, Peterson, & Pardos, 2017). Additionally, a 5-fold cross validation procedure was carried out for tuning "batch_size" and "epoch". The batch size defines the number of samples that will be propagated through the network. The weights are updated after each propagation. The number of epochs is a hyperparameter that defines the number of times that the learning algorithm will work through the entire training dataset. Usually, the model performance increases as the number of epochs increases, but the model begins to overfit when the number of epochs is too large. Therefore, the best epoch number needs to be found. The optimized "batch_size" was 8 and "best epoch" was 46. The final model was trained on all data, with the optimized hyperparameters.

LSTM

The Long Short-Term Memory (LSTM) architecture belongs as part of the family of recurrent neural network architectures. Existing research in the domain of language modeling has found that sequence models based on Long Short-Term Memory networks have strong performance (Sundermeyer, Schlüter, & Ney, 2012), beating prior approaches based on n-grams, hand-crafted features, and "simple" or "vanilla" recurrent neural networks. Utilizing LSTM networks specifically trained on clickstream data has also been used to predict student behaviors in Massively Open Online Courses, to better understand usage patterns as well as to possibly identify useful resources based on the resources similar students have utilized in the past (Tang, Peterson, & Pardos, 2017).

Keras is once again used to implement the LSTM models for this study. All of the hyperparameters in Table 1 apply to our LSTM model implementation as well, except that the number of layers was fixed to 2 for the LSTM approach. Similar to our implementation of the simple RNN model, a 5-fold cross-validation procedure was carried out for hyperparameter tunning on “batch_size” and “epoch”. The optimized “batch_size” was 8 and “best epoch” was 31.

MCNA

A baseline model is named as the “Most Common Next Action” (MCNA). As the name implies, the MCNA model always predicts that the next action will be the most common action that follows the current action, based on the set of training data. This is equivalent to a 2-gram or bigram model, which is equivalent to an *n*-gram model where *n* is set to 2. For this study, the entire available dataset sample is used as the “set of training data” to determine the most common next action for each possible clickstream action.

Statistics of Interest

MAI definition

The Model Agreement Index (MAI) is a straightforward index of how well an examinee’s behaviors align with the trained clickstream behavior model. The index is simply the average probability score of an examinee’s observed actions according to the model’s predictions of their actions. Therefore, MAI is effectively a summarized weighted probability over all actions taken within an individual clickstream.

A clickstream ***c*** can be defined as a list of vectors. Each vector is a representation of a single click taken by an examinee. The dimensionality of each vector is equal to the number of different possible actions in the clickstream data. Each vector is one-hot encoded, meaning that all values of the vector are set to 0, except for one index which is set to 1; this value of 1 corresponds to the action taken at that point in the clickstream.

To calculate MAI for a clickstream ***c***, the corresponding probability from the model output probability distribution for the actual action taken at each timestep is iteratively obtained, summed up, and divided by the length of ***c***.

The MAI formula for a clickstream ***c*** can be described as:

$$MAI_c = \frac{\sum_{s=1}^{S}\sum_{i=1}^{n} t_{si}p_{si}}{S}, \qquad t_{si} = \begin{cases} 1 \text{ if action } i \text{ is the action observed at timestep } s \\ 0 \text{ otherwise} \end{cases} \tag{1}$$

where *S* is the length of the clickstream, *s* represents a single "step" or "timestep" and iterates from 1 through S, *i* is used to correspond to an index used to represent a particular action, *n* is the total number of possible actions and represents the highest possible value of *i*, t_{si} is a truth label at timestep *s* and for action *i* defined as described in formula (1), and p_{si} is the softmax probability from the model for action *i* at timestep *s*.

MAI takes a score range from 0 to 1. Higher scores show stronger agreement between examinee observed behaviors and predicted model actions. Conversely, lower scores mean that the examinee has taken more atypical (and less likely) actions, according to the model's predictions. In general, MAI can be used to identify individual examinee atypical behavior. MAI can also be aggregated for group-level analysis.

Top-1 Accuracy

The prediction accuracy of the prediction models is also evaluated by a top-1 accuracy index. This index evaluates the probability that the observed action is correctly predicted as the most likely action by the prediction model.

$$Top1\ Accuracy_c = \frac{\sum_{s=1}^{S}(predicted\_action_s = observed\_action_s)}{S} \tag{2}$$

Results

Descriptive Statistics

MAI scores and top-1 accuracy are computed for each of the three models, LSTM, RNN, and MCNA. Figure 4 shows the distribution of MAI scores and top-1 accuracy. The density plots for both statistics show the difference between MCNA and LSTM.

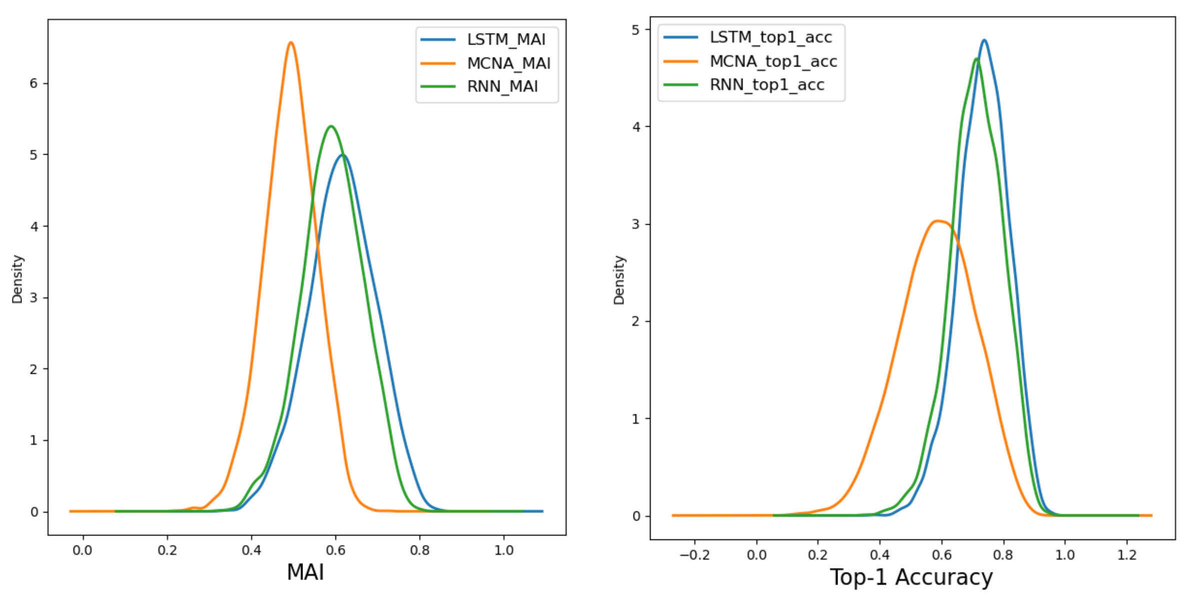

Figure 1 Plot of MAI and Top-1 accuracy distributions

Table 2 Descriptive statistics

	MAI			TOP1_ACC		
	LSTM	Simple RNN	MCNA	LSTM	Simple RNN	MCNA
N count	3934	3934	3934	3934	3934	3934
mean	**0.62**	**0.59**	**0.49**	**0.73**	**0.71**	**0.59**
std	**0.08**	**0.07**	**0.06**	**0.08**	**0.08**	**0.12**
min	0.34	0.32	0.22	0.38	0.35	0.12
25%	0.56	0.55	0.45	0.68	0.66	0.50
50%	0.62	0.59	0.49	0.74	0.71	0.59
75%	0.67	0.64	0.54	0.79	0.77	0.67
max	0.84	0.80	0.72	0.95	0.94	0.89

Table 2 and Figure 1 show the descriptive statistics and distribution curves of the calculated MAI scores by different methods. In summary, the MAI scores calculated by LSTM and simple RNN are higher than those calculated by MCNA, with the LSTM having the highest mean MAI scores. LSTM shows the strongest prediction accuracy among the three models. The average top-1 prediction accuracy of LSTM is 0.73, which is higher by 0.14 than that of MCNA approach.

Model Comparison

Table 3 Comparison of MAIs by LSTM, MCNA, and RNN

		LSTM vs RNN	LSTM vs MCNA	RNN vs MCNA
Absolute Difference of MAI	Mean (S.D.)	.03(.02)	.12(.05)	.10(.04)
	Min	.00	.00	.00
	Max	.16	.41	.39
Correlation Coefficient		.97	.79	.84

In Table 3, some statistics for comparing the MAI by different methods are presented. The first row shows the mean and standard deviation of MAI difference between each pair of methods. The two rows below show the minimum and maximum MAI difference, while the last row shows the Pearson's correlation coefficient between each pair of methods. The average difference between LSTM MAI and RNN MAI is small (0.03), with a standard deviation of 0.02. The MAI values based on these two methods are also highly correlated with a correlation coefficient of 0.97. On the contrary, the average difference between LSTM MAI and MCNA MAI is relatively high (0.12), with a standard deviation of 0.05. The maximum difference is as large as 0.41. The correlation coefficient is moderate: 0.79.

Table 4 The confusion matrix for comparing LSTM, RNN and MCNA (TOP 1 ACCURACY)

		MCNA		
		Correct	Incorrect	Total
	Correct	283951(53.8%)	**108991(20.7%)**	392942(74.5%)
LSTM	Incorrect	22735(4.3%)	112017(21.2%)	134752(25.5%)
	Total	306686(58.1%)	221008(41.9%)	
		RNN		
		Correct	Incorrect	Total
	Correct	369048(69.9%)	**23894(4.5%)**	392942(74.5%)
LSTM	Incorrect	11808(2.2%)	122944(23.3%)	134752(25.5%)
	Total	380856(72.2%)	146838(27.8%)	
		MCNA		
		Correct	Incorrect	Total
	Correct	280886(53.2%)	**99970(18.9%)**	380856(72.2%)
RNN	Incorrect	25800(4.9%)	121038(22.9%)	146838(27.8%)
	Total	306686(58.1%)	221008(41.9%)	

Table 4 shows the confusion matrix for comparing prediction accuracy of the three methods. One key result is that of the total 527,694 actions, the LSTM model predicted 86,256 more actions correctly compared to the MCNA model. This shows that the LSTM approach seems to be better at predicting actions more accurately compared to the MCNA model.

Comparisons to Scale Sores

Each test-taker was assigned to take two testing sessions, denoted as Session 1 and Session 2. Based on response patterns from both Session 1 and Session 2 combined, each test-taker was assigned a scale score that ranges between 200 to 400, indicating the math capability of the test-taker. In this study, MAI scores are calculated for Session 1 only. Considering that students submitted the test after each test session, the actions between two test sessions are not a continuous sequence.

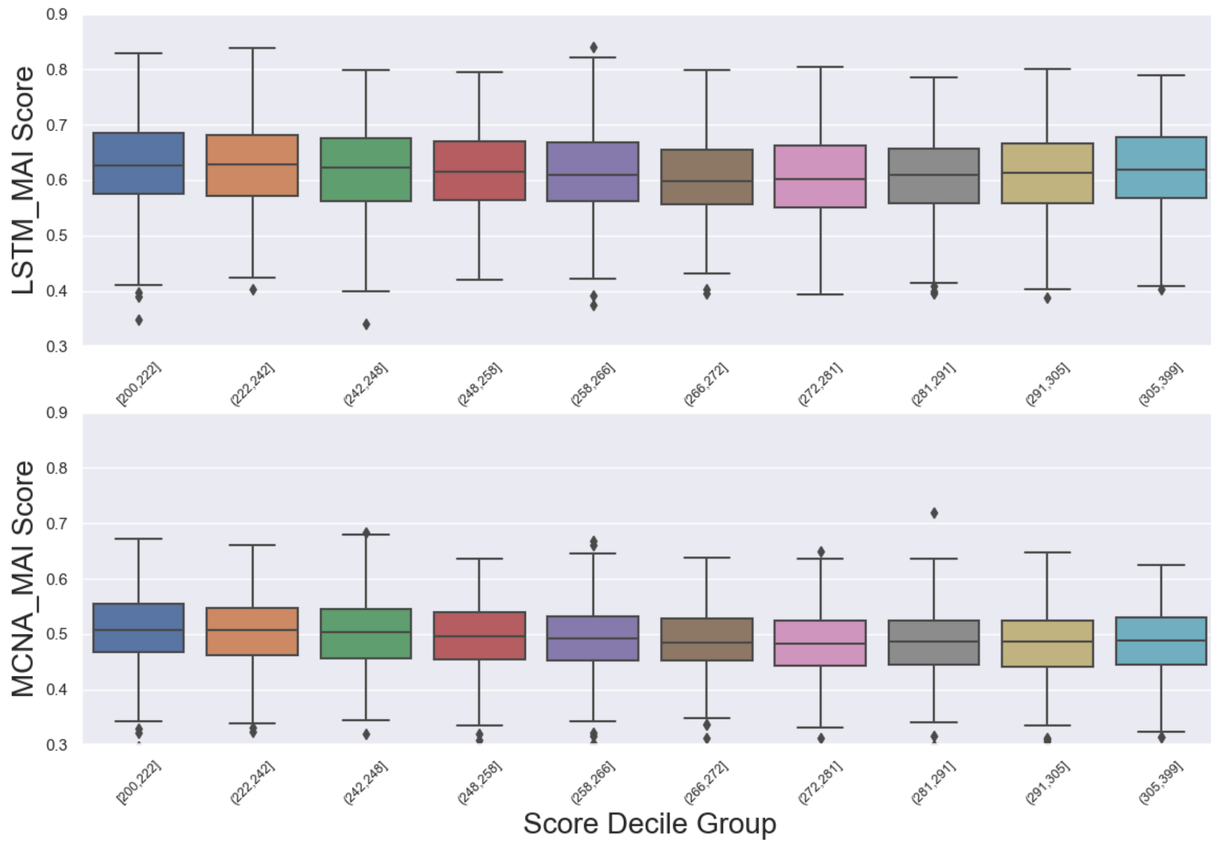

Figure 2 MAI scores against scale score decile groups

Figure 2 plots MAI across the deciles of the scale score distribution. A decile splits the distribution of scale scores into 10 ordered groups, with each decile comprising 10% of the total count of test-takers. The first decile is comprised of the lowest scoring 10% of test-takers, while the last and tenth decile considers the highest scoring 10% of test-takers. The x-axis of the figure shows the range of scores that are included in each decile group. LSTM MAI and MCNA MAI scores are plotted separately. For LSTM MAI results, there appears to be a slightly decreasing trend in median MAI scores up until about the 6th decile group. From the 7th through 10th decile, there is a slightly increasing trend. For MCNA MAI results, the slightly decreasing trend goes from the 1st through the 7th decile, and then there appears to be a slight increase in MAI scores in the 8th decile. These results indicate that the relationship between MAI and performance does not appear to be linear. It is also of note that the inter-quartile ranges of each box plot span a relatively wide range, indicating that there is not necessarily a strong or obvious relationship between MAI and scale score, other than the slight dip observed in the distributions from both test sessions.

Comparing MAI to Traditional Aberrance Detection Statistics

N2 and NC2 (Bishop & Egan, 2017) are two common aberrance indices (Ranger, Schmidt, & Wolgast, 2020) that are relatively straightforward to compute. N2 indicates the number of items on which an examinee changes his/her response at least once. NC2 indicates the number of items on which a test-taker changes his/her response from wrong to right at the last attempt. Other aberrance indices focus on response-time analysis. Based on the lognormal model for response times (van der Linden & Guo, 2008), Li et al. (2018) introduced the statistical index Z_S. Z_S is an item-level index. For this study, we focus on using only the *last* response time recorded by each examinee for each item, disregarding response times for any answer choices other than what ends up as the final response selection for the examinee. High values of Z_S^2 identify where an examinee's response time is unusually quick or unusually slow based on the response times from the entire population of examinees for that item. Z_S is adjusted by an examinee's overall speed for the entire test session. The extent of aberrance of an examinee's response time pattern for the entire test is represented by the average of Z_S^2 across all items.

Table 5 Correlation Coefficients Between MAI scores and Traditional Aberrance Detection Indices

	LSTM		Simple RNN		MCNA	
	MAI_Score	Top1_Acc	MAI_Score	Top1_Acc	MAI_Score	Top1_Acc
N2	-0.28	-0.32	-0.28	-0.32	-0.18	-0.17
NC2	-0.23	-0.26	-0.24	-0.26	-0.18	-0.19
$\boldsymbol{Average\_Z_S^2}$	-0.20	-0.21	-0.19	-0.21	-0.11	-0.11

From Table 5, among the traditional aberrance detection indices, both N2 and NC2 have a weak negative correlation with MAI by LSTM/simple RNN. The correlation coefficients are even smaller for the MAI by MCNA. The correlation between N2 and MAI scores is the highest among the tested statistics; this could be somewhat expected given that both N2 and the current MAI approach do not consider response correctness or response times, while the other models do. The negative correlation shows that, on average, examinees who change answers more frequently have lower MAI scores.

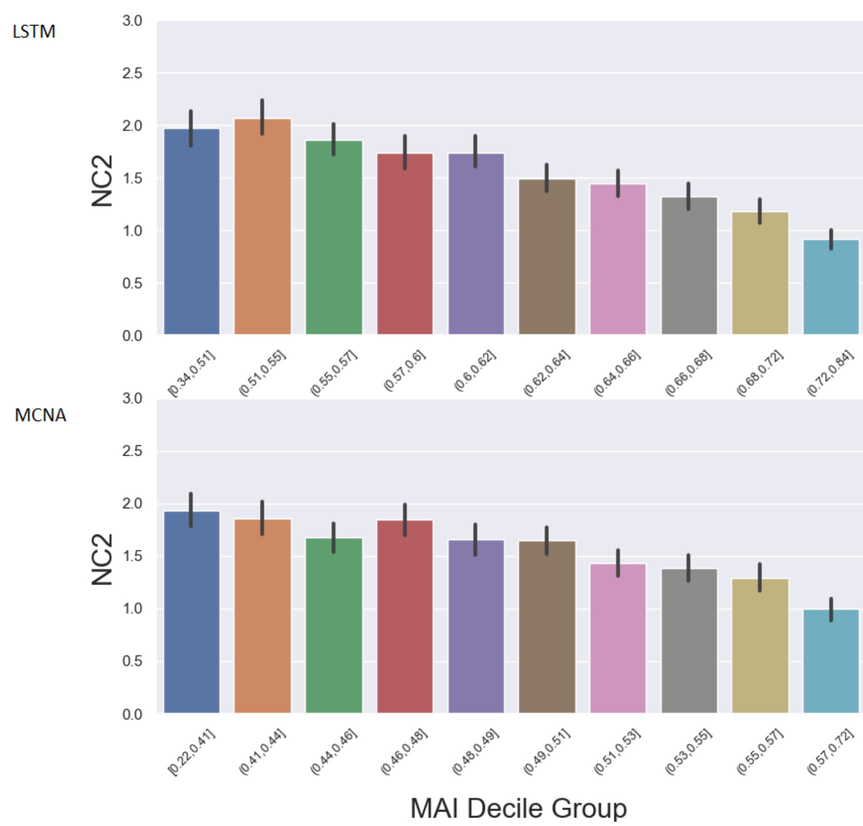

Figure 3 NC2 Index Across LSTM MAI Deciles/ MCNA MAI Deciles

Figure 3 plots the average NC2 value across the 10 deciles of MAI scores. There is a downward trend for both LSTM MAI and MCNA MAI. This trend shows that lower MAI scores tended to have higher NC2 values across the entire distribution of MAI scores. In interpretive terms, this means that clickstreams that were identified as relatively more atypical by their MAI values tended to also be relatively more aberrant according to their NC2 values. For LSTM MAI scores, the decreasing of NC2 is more obvious across the MAI score deciles.

The correlation coefficient between MAI scores and the response time index Z_s^2 is slightly negative. Examinees who have higher response time aberrance on their last attempt on an item tended to have slightly lower MAI values. The current calculation of MAI does not incorporate response time or timing between actions. In future work, if timings were to be included as part of the MAI computation, correlations with aberrance indices that are related to response times could increase.

What actions are commonly observed in Low-MAI and High-MAI clickstreams?

We define “Low MAI” to include MAI values that are lower than 2 standard deviations below the mean. We define “High MAI” to include MAI values higher than 2 standard deviations above the mean. With this definition, among the 3934 clickstreams, 104 clickstreams are in the “Low MAI” group, while 67 clickstreams are in the “High MAI” group. The “Low MAI” clickstreams contain 8831 actions in total and the “High MAI” clickstreams contain 7664 actions in total.

Table 6 breaks down the distribution of the 8831 observed actions from the Low MAI group by also considering what the most likely action predicted by the behavior model was when that observed action occurred. For example, row 1 describes the % of all observed actions where the observed action was a NAVIGATION_ITEM_NEXT and the predicted action at that point in time was also a NAVIGATION_ITEM_NEXT. On the other hand, row 8 depicts the % of all observed actions where the observed action was a NAVIGATION_ITEM_NEXT but the prediction model at that point in time predicted a different action, specifically ITEM_MULTIPLE_CHOICE_ANSWER. Table 6 shows the top 20 most frequent observed/prediction action pairs, sorted in descending order in terms of the frequency of each “observed action” and “predicted action” pair. Any row highlighted in **bold** shows a mismatching prediction pair. Additionally, the last column states whether the observed and predicted action are a match for that row.

Table 7 shows the same information but for the distribution of the 7664 actions from the High MAI group.

Table 6 Percentages of observed/predicted action pairs in "Low MAI" group

Row	Observed Action	Predicted Action by LSTM	%	Label
1	NAVIGATION_ITEM_NEXT	NAVIGATION_ITEM_NEXT	12.3%	Match
2	ITEM_MULTIPLE_CHOICE_ANSWER	ITEM_MULTIPLE_CHOICE_ANSWER	9.7%	Match
3	**ITEM_MULTIPLE_CHOICE_ANSWER**	**NAVIGATION_ITEM_NEXT**	**4.5%**	
4	TOOL_CALCULATOR_OPEN	TOOL_CALCULATOR_OPEN	4.2%	Match
5	TOOL_ANSWER_MASKING_TOGGLE	TOOL_ANSWER_MASKING_TOGGLE	3.5%	Match
6	NAVIGATION_REVIEW_PANEL_CLOSE	NAVIGATION_REVIEW_PANEL_CLOSE	3.4%	Match
7	**TOOL_CALCULATOR_TOGGLE**	**ITEM_MULTIPLE_CHOICE_ANSWER**	**2.6%**	
8	**NAVIGATION_ITEM_NEXT**	**ITEM_MULTIPLE_CHOICE_ANSWER**	**2.5%**	
9	TOOL_CALCULATOR_CLOSE	TOOL_CALCULATOR_CLOSE	1.9%	Match
10	**TOOL_ANSWER_MASKING_TOGGLE**	**ITEM_MULTIPLE_CHOICE_ANSWER**	**1.9%**	
11	TOOL_SKETCH_SELECT	TOOL_SKETCH_SELECT	1.9%	Match
12	**TOOL_CALCULATOR_CLOSE**	**ITEM_MULTIPLE_CHOICE_ANSWER**	**1.6%**	
13	NAVIGATION_ITEM_BACK	NAVIGATION_ITEM_BACK	1.5%	Match
14	**TOOL_CALCULATOR_TOGGLE**	**TOOL_CALCULATOR_OPEN**	**1.5%**	
15	NAVIGATION_ITEM_JUMP	NAVIGATION_ITEM_JUMP	1.5%	Match
16	**TOOL_ANSWER_MASKING_TOGGLE**	**NAVIGATION_ITEM_NEXT**	**1.3%**	
17	TOOL_CALCULATOR_TOGGLE	TOOL_CALCULATOR_TOGGLE	1.2%	Match
18	**NAVIGATION_REVIEW_PANEL_OPEN**	**NAVIGATION_ITEM_NEXT**	**1.2%**	
19	NAVIGATION_TURN_IN_COMMIT	NAVIGATION_TURN_IN_COMMIT	1.2%	Match
20	**NAVIGATION_REVIEW_PANEL_OPEN**	**ITEM_MULTIPLE_CHOICE_ANSWER**	**1.1%**	

Table 7 Percentages of observed/predicted action pairs in "High MAI" group

Observed Action	Predicted Action by LSTM	Percent	Label
NAVIGATION_ITEM_NEXT	NAVIGATION_ITEM_NEXT	25.5%	Match
ITEM_MULTIPLE_CHOICE_ANSWER	ITEM_MULTIPLE_CHOICE_ANSWER	24.3%	Match
ITEM_DRAG_BOX_DRAG_END	ITEM_DRAG_BOX_DRAG_END	5.5%	Match
ITEM_DRAG_BOX_DRAG_START	ITEM_DRAG_BOX_DRAG_START	5.5%	Match
ITEM_TILE_BOX_DRAG_END	ITEM_TILE_BOX_DRAG_END	4.1%	Match
ITEM_TILE_BOX_DRAG_START	ITEM_TILE_BOX_DRAG_START	4.0%	Match
TOOL_ANSWER_MASKING_TOGGLE	TOOL_ANSWER_MASKING_TOGGLE	4.0%	Match
ITEM_SELECT_DROP_DOWN_select	ITEM_SELECT_DROP_DOWN_select	2.4%	Match
NAVIGATION_REVIEW_PANEL_CLOSE	NAVIGATION_REVIEW_PANEL_CLOSE	2.0%	Match
ITEM_MULTIPLE_CHOICE_ANSWER	**NAVIGATION_ITEM_NEXT**	**1.6%**	

NAVIGATION_ACCESS_CODE_SUBMIT	NAVIGATION_ACCESS_CODE_SUBMIT	1.5%	Match
ITEM_BOOKMARK_OFF	ITEM_BOOKMARK_OFF	1.0%	Match
NAVIGATION_REVIEW_PANEL_OPEN	NAVIGATION_REVIEW_PANEL_OPEN	1.0%	Match
ITEM_BOOKMARK_ON	ITEM_BOOKMARK_ON	1.0%	Match
NAVIGATION_PROFILE_CHOOSE	NAVIGATION_PROFILE_CHOOSE	0.9%	Match
NAVIGATION_PROFILE_LOGIN	NAVIGATION_PROFILE_LOGIN	0.9%	Match
NAVIGATION_TURN_IN_COMMIT	NAVIGATION_TURN_IN_COMMIT	0.9%	Match
NAVIGATION_TURN_IN_START	NAVIGATION_TURN_IN_START	0.9%	Match
NAVIGATION_ITEM_NEXT	**ITEM_TILE_BOX_DRAG_START**	**0.8%**	
NAVIGATION_ITEM_NEXT	**ITEM_DRAG_BOX_DRAG_START**	**0.8%**	

Table 6 and Table 7 show that more mismatched observed/prediction action pairs exist for the low MAI group than the high MAI group. Among the 20 action pairs, 9 in the low MAI group are mismatched pairs, while only 3 in the high MAI group are mismatched pairs. The percents of mismatched observed/prediction action pairs are also much higher in the low MAI group. The most common mismatched pair in both the low and high MAI groups is the same: when the observed action is "ITEM_MULTIPLE_CHOICE_ANSWER", the predicted action is "NAVIGATION_ITEM_NEXT". The percentage of this pair is 4.5% for the low MAI group, while it is only 1.6% for the high MAI group. Additionally, the percents of matched observed/prediction action pairs are much higher in the high MAI group. For example, two matched events, "NAVIGATION_ITEM_NEXT" and "ITEM_MULTIPLE_CHOICE_ANSWER", have the highest probabilities in both the low MAI and high MAI groups. However, in the high MAI group, the percentages of the two most matched action pairs took approximately 50% of the total action pairs, while their percentages only summed up to 22% in the low MAI group.

The low MAI group contains several mismatched action pairs related to tool usage, which is not observed in the high MAI group. Specifically, the action of "TOOL_CALCULATOR_TOGGLE" was frequently observed when the predicted action is "ITEM_MULTIPLE_CHOICE_ANSWER". In addition, "TOOL_ANSWER_MASKING_TOGGLE", "TOOL_CALCULATOR_CLOSE", "TOOL_ANSWER_MASKING_TOGGLE" are also among the identified atypical clickstream actions in the low MAI group. These atypical clickstream actions might indicate test-takers' misuse or misunderstanding of the tools. Clickstream examples will be introduced in the following section to further explain in what conditions test-takers might use the tools in unexpected ways.

It can also be noticed that the low MAI group and high MAI group are different regarding how test-takers use the review panels. In the high MAI group, the action of "NAVIGATION_REVIEW_PANEL_OPEN" seems to be matched with the prediction. Test-takers use the review panel as predicted. However, in the low MAI group, the action of "NAVIGATION_REVIEW_PANEL_OPEN" is often not matched with the prediction. The test-takers seem to be more likely to open the review panel when the predicted action is "NAVIGATION_ITEM_NEXT" or "ITEM_MULTIPLE_CHOICE_ANSWER".

Table 12 in the appendix shows the full list of mismatched events in the low MAI group.

Examples

In this section, three types of clickstreams are analyzed: 1) clickstream with low MAI by LSTM; 2) clickstream with high MAI by LSTM; 3) clickstreams with large differences on MAI scores between LSTM and MCNA.

Clickstream Example with low MAI by LSTM

Table 8 shows the list of actions (ordered sequentially) for an example clickstream that obtained a low MAI score in this dataset. The corresponding predicted probabilities by LSTM are listed in the last column. In this clickstream, a few peculiar conclusions can be obtained. Firstly, the test-taker starts the test with many actions on using the tools on the first item. This a very rare clickstream pattern. It seems that the test-taker intends to examine the functionality of each tool carefully before reading and answering any test questions. Additionally, the test-taker often toggles the tools during testing, which is also a relatively uncommon task. Thirdly, the end of this clickstream is "ALERT_INACTIVITY_EXIT" event instead of "ALERT_PROFILE_EXIT", meaning that the test-taker didn't exit the exam appropriately.

*Table 8 List of actions and predicted probabilities for clickstream with **low MAI** by LSTM*

Step	Observed Action	Predicted Probability by LSTM
1	NAVIGATION_PROFILE_LOGIN	0.94
2	NAVIGATION_PROFILE_CHOOSE	0.90
3	NAVIGATION_ACCESS_CODE_SUBMIT	0.93
4	NAVIGATION_DIRECTIONS_CONTINUE	0.84
5	TOOL_TEXT_HIGHLIGHT_TOGGLE	0.01
6	TOOL_TEXT_HIGHLIGHT_SELECTED	0.29
7	TOOL_TEXT_HIGHLIGHT_CANCEL	0.24
8	TOOL_TEXT_HIGHLIGHT_CANCEL	0.35
9	TOOL_TEXT_HIGHLIGHT_CANCEL	0.57
10	TOOL_TEXT_HIGHLIGHT_CANCEL	0.54
11	TOOL_TEXT_HIGHLIGHT_TOGGLE	0.23
12	TOOL_SKETCH_SELECT	0.14
13	TOOL_SKETCH_OPEN	0.88
14	TOOL_SKETCH_SELECT	0.88
15	TOOL_SKETCH_SELECT	0.51
16	TOOL_SKETCH_CLOSE	0.59
17	TOOL_TEXT_HIGHLIGHT_TOGGLE	0.56
18	TOOL_TEXT_HIGHLIGHT_CANCEL_ALL	0.40
19	TOOL_TEXT_HIGHLIGHT_CANCEL_ALL	0.22
20	TOOL_TEXT_HIGHLIGHT_CANCEL_ALL	0.31
21	TOOL_TEXT_HIGHLIGHT_CANCEL_ALL	0.42
22	TOOL_SKETCH_SELECT	0.13
23	TOOL_SKETCH_OPEN	0.99
24	TOOL_SKETCH_SELECT	0.86
25	TOOL_SKETCH_SELECT	0.22
26	TOOL_SKETCH_SELECT	0.20
27	TOOL_SKETCH_CLOSE	0.53
28	TOOL_REFERENCES_TOGGLE	0.10
29	TOOL_REFERENCES_TOGGLE	0.21
30	TOOL_REFERENCES_TOGGLE	0.46
31	TOOL_REFERENCES_OPEN	0.47
32	TOOL_REFERENCES_CLOSE	0.74
33	ITEM_STIMULUS_TOGGLE	0.17
34	ITEM_STIMULUS_TOGGLE	0.97
35	ITEM_MULTIPLE_CHOICE_ANSWER	0.39
36	TOOL_GUIDELINE_OPEN	0.01
37	TOOL_GUIDELINE_CLOSE	0.72
38	TOOL_GUIDELINE_OPEN	0.10
39	TOOL_GUIDELINE_CLOSE	0.96
40	TOOL_GUIDELINE_OPEN	0.25
41	TOOL_GUIDELINE_CLOSE	0.99
42	NAVIGATION_ITEM_NEXT	0.15

43	ITEM_MULTIPLE_CHOICE_ANSWER	0.46
44	ITEM_MULTIPLE_CHOICE_ANSWER	0.23
45	TOOL_REFERENCES_CLOSE	0.00
46	NAVIGATION_ITEM_NEXT	0.20
47	TOOL_ANSWER_MASKING_TOGGLE	0.02
48	TOOL_ANSWER_MASKING_TOGGLE	0.79
49	ITEM_MULTIPLE_CHOICE_ANSWER	0.59
50	NAVIGATION_ITEM_NEXT	0.58
51	TOOL_REFERENCES_TOGGLE	0.02
52	ITEM_MULTIPLE_CHOICE_ANSWER	0.01
53	NAVIGATION_ITEM_NEXT	0.42
54	NAVIGATION_ITEM_NEXT	0.04
55	TOOL_ANSWER_MASKING_TOGGLE	0.11
56	TOOL_REFERENCES_TOGGLE	0.01
57	TOOL_REFERENCES_OPEN	0.70
58	TOOL_REFERENCES_CLOSE	0.57
59	TOOL_ANSWER_MASKING_TOGGLE	0.15
60	ITEM_SELECT_DROP_DOWN_select	0.01
61	ITEM_MULTIPLE_CHOICE_ANSWER	0.09
62	NAVIGATION_ITEM_NEXT	0.47
63	TOOL_ANSWER_MASKING_TOGGLE	0.17
64	TOOL_ANSWER_MASKING_TOGGLE	0.75
65	ITEM_MULTIPLE_CHOICE_ANSWER	0.63
66	NAVIGATION_ITEM_NEXT	0.58
67	ITEM_MULTIPLE_CHOICE_ANSWER	0.49
68	NAVIGATION_ITEM_NEXT	0.66
69	ITEM_MULTIPLE_CHOICE_ANSWER	0.50
70	NAVIGATION_ITEM_NEXT	0.67
71	TOOL_ANSWER_MASKING_TOGGLE	0.12
72	ITEM_MULTIPLE_CHOICE_ANSWER	0.17
73	NAVIGATION_ITEM_NEXT	0.55
74	ITEM_MULTIPLE_CHOICE_ANSWER	0.44
75	NAVIGATION_ITEM_NEXT	0.61
76	ITEM_MULTIPLE_CHOICE_ANSWER	0.43
77	ITEM_BOOKMARK_OFF	0.03
78	ITEM_SELECT_DROP_DOWN_select	0.00
79	ITEM_SELECT_DROP_DOWN_select	0.63
80	TOOL_REFERENCES_TOGGLE	0.00
81	TOOL_REFERENCES_OPEN	0.55
82	ITEM_MULTIPLE_CHOICE_ANSWER	0.15
83	TOOL_REFERENCES_CLOSE	0.27
84	ITEM_BOOKMARK_OFF	0.22
85	NAVIGATION_ITEM_NEXT	0.11
86	NAVIGATION_ITEM_NEXT	0.19
87	ITEM_MULTIPLE_CHOICE_ANSWER	0.15
88	ITEM_BOOKMARK_OFF	0.10
89	NAVIGATION_ITEM_NEXT	0.30
90	NAVIGATION_ITEM_NEXT	0.42

91	NAVIGATION_REVIEW_PANEL_OPEN	0.12
92	NAVIGATION_TURN_IN_START	0.52
93	NAVIGATION_REVIEW_PANEL_CLOSE	0.98
94	NAVIGATION_TURN_IN_COMMIT	1.00
95	ALERT_INACTIVITY_EXIT	0.09
	End Token	0.73
	MAI	**0.41**

Clickstream Example with high MAI by LSTM

Table 9 shows the list of actions (ordered sequentially) and their corresponding predicted probabilities by LSTM for an example clickstream with a high MAI score. This clickstream consists of two main actions: navigating to the next item and answering the items. On step 79, when the test-taker suddenly opened the review panel, the action of "NAVIGATION_REVIEW_PANEL_OPEN" has a low predicted probability. However, when it appears on step 92, where the test is almost finished, the predicted probability is very high.

*Table 9 List of actions and predicted probabilities for clickstream with **high MAI** by LSTM*

Step	Observed Action	Predicted Probability by LSTM
1	NAVIGATION_PROFILE_LOGIN	0.94
2	NAVIGATION_PROFILE_CHOOSE	0.90
3	NAVIGATION_ACCESS_CODE_SUBMIT	0.93
4	NAVIGATION_DIRECTIONS_CONTINUE	0.84
5	ITEM_MULTIPLE_CHOICE_ANSWER	0.38
6	NAVIGATION_ITEM_NEXT	0.51
7	ITEM_MULTIPLE_CHOICE_ANSWER	0.31
8	NAVIGATION_ITEM_NEXT	0.77
9	ITEM_DRAG_BOX_DRAG_START	0.82
10	ITEM_DRAG_BOX_DRAG_END	1.00
11	ITEM_DRAG_BOX_DRAG_START	0.95
12	ITEM_DRAG_BOX_DRAG_END	1.00
13	ITEM_DRAG_BOX_DRAG_START	0.96
14	ITEM_DRAG_BOX_DRAG_END	1.00
15	ITEM_DRAG_BOX_DRAG_START	0.96
16	ITEM_DRAG_BOX_DRAG_END	1.00
17	ITEM_DRAG_BOX_DRAG_START	0.64
18	ITEM_DRAG_BOX_DRAG_END	1.00
19	NAVIGATION_ITEM_NEXT	0.35
20	ITEM_MULTIPLE_CHOICE_ANSWER	0.70
21	NAVIGATION_ITEM_NEXT	0.83
22	ITEM_MULTIPLE_CHOICE_ANSWER	0.84

23	NAVIGATION_ITEM_NEXT	0.79
24	ITEM_MULTIPLE_CHOICE_ANSWER	0.82
25	NAVIGATION_ITEM_NEXT	0.84
26	ITEM_MULTIPLE_CHOICE_ANSWER	0.84
27	NAVIGATION_ITEM_NEXT	0.84
28	ITEM_MULTIPLE_CHOICE_ANSWER	0.87
29	NAVIGATION_ITEM_NEXT	0.83
30	ITEM_MULTIPLE_CHOICE_ANSWER	0.88
31	NAVIGATION_ITEM_NEXT	0.83
32	ITEM_MULTIPLE_CHOICE_ANSWER	0.88
33	NAVIGATION_ITEM_NEXT	0.84
34	ITEM_MULTIPLE_CHOICE_ANSWER	0.88
35	NAVIGATION_ITEM_NEXT	0.84
36	ITEM_MULTIPLE_CHOICE_ANSWER	0.88
37	NAVIGATION_ITEM_NEXT	0.85
38	ITEM_MULTIPLE_CHOICE_ANSWER	0.88
39	NAVIGATION_ITEM_NEXT	0.85
40	ITEM_MULTIPLE_CHOICE_ANSWER	0.88
41	NAVIGATION_ITEM_NEXT	0.86
42	ITEM_MULTIPLE_CHOICE_ANSWER	0.89
43	NAVIGATION_ITEM_NEXT	0.87
44	ITEM_MULTIPLE_CHOICE_ANSWER	0.89
45	NAVIGATION_ITEM_NEXT	0.89
46	ITEM_MULTIPLE_CHOICE_ANSWER	0.77
47	NAVIGATION_ITEM_NEXT	0.91
48	ITEM_SELECT_DROP_DOWN_select	0.85
49	ITEM_SELECT_DROP_DOWN_select	0.98
50	ITEM_SELECT_DROP_DOWN_select	0.99
51	NAVIGATION_ITEM_NEXT	0.58
52	ITEM_MULTIPLE_CHOICE_ANSWER	0.91
53	NAVIGATION_ITEM_NEXT	0.88
54	ITEM_MULTIPLE_CHOICE_ANSWER	0.92
55	NAVIGATION_ITEM_NEXT	0.86
56	ITEM_MULTIPLE_CHOICE_ANSWER	0.92
57	NAVIGATION_ITEM_NEXT	0.88
58	ITEM_MULTIPLE_CHOICE_ANSWER	0.92
59	NAVIGATION_ITEM_NEXT	0.89
60	ITEM_MULTIPLE_CHOICE_ANSWER	0.92
61	NAVIGATION_ITEM_NEXT	0.90
62	ITEM_MULTIPLE_CHOICE_ANSWER	0.92
63	NAVIGATION_ITEM_NEXT	0.90
64	ITEM_MULTIPLE_CHOICE_ANSWER	0.93
65	NAVIGATION_ITEM_NEXT	0.88
66	ITEM_MULTIPLE_CHOICE_ANSWER	0.84
67	NAVIGATION_ITEM_NEXT	0.87
68	ITEM_TILE_BOX_DRAG_START	0.90
69	ITEM_TILE_BOX_DRAG_END	1.00
70	ITEM_TILE_BOX_DRAG_START	0.98

71	ITEM_TILE_BOX_DRAG_END	1.00
72	ITEM_TILE_BOX_DRAG_START	0.97
73	ITEM_TILE_BOX_DRAG_END	1.00
74	ITEM_TILE_BOX_DRAG_START	0.66
75	ITEM_TILE_BOX_DRAG_END	1.00
76	ITEM_TILE_BOX_DRAG_START	0.72
77	ITEM_TILE_BOX_DRAG_END	1.00
78	NAVIGATION_ITEM_NEXT	0.27
79	NAVIGATION_REVIEW_PANEL_OPEN	0.03
80	NAVIGATION_REVIEW_PANEL_CLOSE	0.99
81	ITEM_MULTIPLE_CHOICE_ANSWER	0.84
82	NAVIGATION_ITEM_NEXT	0.89
83	ITEM_MULTIPLE_CHOICE_ANSWER	0.92
84	NAVIGATION_ITEM_NEXT	0.83
85	ITEM_MULTIPLE_CHOICE_ANSWER	0.93
86	NAVIGATION_ITEM_NEXT	0.87
87	ITEM_MULTIPLE_CHOICE_ANSWER	0.92
88	NAVIGATION_ITEM_NEXT	0.89
89	ITEM_MULTIPLE_CHOICE_ANSWER	0.92
90	NAVIGATION_ITEM_NEXT	0.86
91	ITEM_MULTIPLE_CHOICE_ANSWER	0.92
92	NAVIGATION_REVIEW_PANEL_OPEN	0.90
93	NAVIGATION_TURN_IN_START	0.84
94	NAVIGATION_REVIEW_PANEL_CLOSE	0.98
95	NAVIGATION_TURN_IN_COMMIT	1.00
96	ALERT_PROFILE_EXIT	0.80
	End Token	0.78
	MAI	**0.85**

Clickstream examples with large differences between LSTM MAI and MCNA MAI

Both the LSTM MAI and the MCNA MAI approach could potentially be useful as predictive behavior models. Both approaches might have substantial overlap in their predictions of actions; however, it is clear from the statistical results that differences exist. In this section, two clickstream examples are shown, whereby the MAI values from the LSTM and MCNA models differed substantially. This analysis can help determine the kinds of behavior patterns that the LSTM MAI approach can more successfully model compared to the MCNA approach.

The first example contains repeated actions of “Navigation Item Back” and “Navigation Item Next”. In the LSTM model, the predicted probability of these actions is much higher than that of the

MCNA model. The repeated actions might indicate that the test-taker is reviewing the items back and forth. This is a common test-taking strategy, where students review multiple items in a row without changing their answers; however, not every student will use this strategy. The trained LSTM has learned to be able to predict this type of pattern when certain actions are repeated successively. On the contrary, MCNA only assigned a fixed low probability to all the repeated actions, causing the MCNA model to assign a low probability to this behavior pattern. The second clickstream example shows the difference between LSTM and MCNA for a clickstream where the actions of “ITEM BOOKMARK ON” and “ITEM BOOKMARK OFF” occur iteratively. This behavior is somewhat odd, as there’s no practical reason for a student to want to engage in this behavior, but when a test-taker starts to repeat this behavior, it’s more likely for this cycle of behaviors to continue, and the LSTM model has learned to better predict these cyclical behaviors. Perhaps this is an interesting discussion point, whereby the MCNA result could be sometimes “preferred” in terms of identifying non-sensical behavior patterns, even if those behavior patterns are observed in practice and predictable by an LSTM approach.

Table 10 Example of clickstream – Repeated actions of “Navigation Item Back” and “Navigation Item Next”

Step	Observed Action	Predicted Probability		
		LSTM	MCNA	RNN
1	NAVIGATION_PROFILE_LOGIN	0.94	0.93	0.93
2	ALERT_PROFILE_EXIT	0.01	0.10	0.01
3	NAVIGATION_PROFILE_LOGIN	0.95	0.15	0.96
4	NAVIGATION_PROFILE_CHOOSE	0.89	0.78	0.95
5	ITEM_MULTIPLE_CHOICE_ANSWER	0.00	0.01	0.01
6	ITEM_MULTIPLE_CHOICE_ANSWER	0.22	0.20	0.24
7	ITEM_MULTIPLE_CHOICE_ANSWER	0.33	0.20	0.52
8	TOOL_ANSWER_MASKING_TOGGLE	0.04	0.02	0.04
9	TOOL_CALCULATOR_TOGGLE	0.00	0.01	0.02
10	TOOL_CALCULATOR_TOGGLE	0.37	0.39	0.48
11	TOOL_CALCULATOR_TOGGLE	0.48	0.39	0.64
12	TOOL_CALCULATOR_OPEN	0.48	0.57	0.26
13	NAVIGATION_ITEM_NEXT	0.09	0.03	0.05
14	ITEM_MULTIPLE_CHOICE_ANSWER	0.24	0.57	0.08
15	NAVIGATION_ITEM_NEXT	0.58	0.70	0.66
16	ITEM_MULTIPLE_CHOICE_ANSWER	0.45	0.57	0.20
17	NAVIGATION_ITEM_NEXT	0.82	0.70	0.80

18	ITEM_MULTIPLE_CHOICE_ANSWER	0.63	0.57	0.30
19	NAVIGATION_ITEM_NEXT	0.79	0.70	0.84
20	ITEM_MULTIPLE_CHOICE_ANSWER	0.63	0.57	0.66
21	NAVIGATION_ITEM_NEXT	0.75	0.70	0.85
22	TOOL_CALCULATOR_TOGGLE	0.07	0.06	0.05
23	TOOL_CALCULATOR_OPEN	0.66	0.57	0.64
24	ITEM_MULTIPLE_CHOICE_ANSWER	0.46	0.38	0.46
25	TOOL_CALCULATOR_CLOSE	0.13	0.01	0.16
26	NAVIGATION_ITEM_NEXT	0.77	0.14	0.80
27	TOOL_CALCULATOR_TOGGLE	0.12	0.06	0.22
28	TOOL_CALCULATOR_OPEN	0.78	0.57	0.76
29	ITEM_MULTIPLE_CHOICE_ANSWER	0.51	0.38	0.51
30	TOOL_CALCULATOR_CLOSE	0.13	0.01	0.19
31	**NAVIGATION_ITEM_BACK**	0.01	0.03	0.01
32	**NAVIGATION_ITEM_BACK**	0.28	0.27	0.07
33	**NAVIGATION_ITEM_BACK**	0.69	0.27	0.45
34	**NAVIGATION_ITEM_BACK**	0.83	0.27	0.39
35	**NAVIGATION_ITEM_BACK**	0.80	0.27	0.55
36	**NAVIGATION_ITEM_BACK**	0.75	0.27	0.48
37	**NAVIGATION_ITEM_BACK**	0.77	0.27	0.71
38	**NAVIGATION_ITEM_BACK**	0.82	0.27	0.73
39	**NAVIGATION_ITEM_BACK**	0.83	0.27	0.79
40	**NAVIGATION_ITEM_BACK**	0.84	0.27	0.79
41	**NAVIGATION_ITEM_BACK**	0.85	0.27	0.83
42	**NAVIGATION_ITEM_BACK**	0.85	0.27	0.83
43	**NAVIGATION_ITEM_BACK**	0.86	0.27	0.85
44	**NAVIGATION_ITEM_BACK**	0.87	0.27	0.83
45	**NAVIGATION_ITEM_BACK**	0.88	0.27	0.84
46	**NAVIGATION_ITEM_BACK**	0.89	0.27	0.83
47	**NAVIGATION_ITEM_BACK**	0.90	0.27	0.85
48	**NAVIGATION_ITEM_BACK**	0.90	0.27	0.85
49	**NAVIGATION_ITEM_BACK**	0.91	0.27	0.85
50	**NAVIGATION_ITEM_BACK**	0.91	0.27	0.86
51	**NAVIGATION_ITEM_BACK**	0.91	0.27	0.86
52	**NAVIGATION_ITEM_BACK**	0.91	0.27	0.86
53	**NAVIGATION_ITEM_BACK**	0.91	0.27	0.86
54	**NAVIGATION_ITEM_BACK**	0.91	0.27	0.86
55	TOOL_CALCULATOR_TOGGLE	0.02	0.03	0.01
56	TOOL_CALCULATOR_TOGGLE	0.17	0.39	0.45
57	TOOL_CALCULATOR_OPEN	0.82	0.57	0.50
58	ITEM_MULTIPLE_CHOICE_ANSWER	0.31	0.38	0.09
59	TOOL_CALCULATOR_CLOSE	0.28	0.01	0.27
60	NAVIGATION_ITEM_BACK	0.50	0.03	0.04
61	NAVIGATION_ITEM_BACK	0.43	0.27	0.07
62	NAVIGATION_ITEM_BACK	0.77	0.27	0.49
63	NAVIGATION_ITEM_BACK	0.84	0.27	0.46

64	NAVIGATION_ITEM_BACK	0.89	0.27	0.59
65	**NAVIGATION_ITEM_NEXT**	0.04	0.28	0.11
66	**NAVIGATION_ITEM_NEXT**	0.65	0.11	0.54
67	**NAVIGATION_ITEM_NEXT**	0.90	0.11	0.60
68	**NAVIGATION_ITEM_NEXT**	0.91	0.11	0.60
69	**NAVIGATION_ITEM_NEXT**	0.88	0.11	0.61
70	**NAVIGATION_ITEM_NEXT**	0.79	0.11	0.72
71	**NAVIGATION_ITEM_NEXT**	0.71	0.11	0.80
72	**NAVIGATION_ITEM_NEXT**	0.68	0.11	0.84
73	**NAVIGATION_ITEM_NEXT**	0.68	0.11	0.87
74	**NAVIGATION_ITEM_NEXT**	0.71	0.11	0.88
75	**NAVIGATION_ITEM_NEXT**	0.77	0.11	0.89
76	**NAVIGATION_ITEM_NEXT**	0.79	0.11	0.89
77	**NAVIGATION_ITEM_NEXT**	0.79	0.11	0.90
78	**NAVIGATION_ITEM_NEXT**	0.80	0.11	0.90
79	**NAVIGATION_ITEM_NEXT**	0.82	0.11	0.90
80	**NAVIGATION_ITEM_NEXT**	0.83	0.11	0.90
81	**NAVIGATION_ITEM_NEXT**	0.84	0.11	0.90
82	**NAVIGATION_ITEM_NEXT**	0.85	0.11	0.90
83	**NAVIGATION_ITEM_NEXT**	0.86	0.11	0.90
84	**NAVIGATION_ITEM_NEXT**	0.86	0.11	0.90
85	**NAVIGATION_ITEM_NEXT**	0.86	0.11	0.90
86	**NAVIGATION_ITEM_NEXT**	0.86	0.11	0.90
87	**NAVIGATION_ITEM_NEXT**	0.86	0.11	0.90
88	**NAVIGATION_ITEM_NEXT**	0.85	0.11	0.90
89	**NAVIGATION_ITEM_NEXT**	0.85	0.11	0.90
90	**NAVIGATION_ITEM_NEXT**	0.85	0.11	0.90
91	**NAVIGATION_ITEM_NEXT**	0.85	0.11	0.90
92	**NAVIGATION_ITEM_NEXT**	0.84	0.11	0.90
93	**NAVIGATION_ITEM_NEXT**	0.84	0.11	0.90
94	NAVIGATION_REVIEW_PANEL_OPEN	0.10	0.04	0.06
95	NAVIGATION_TURN_IN_START	0.72	0.23	0.60
96	NAVIGATION_REVIEW_PANEL_CLOSE	0.98	0.98	0.95
97	NAVIGATION_TURN_IN_COMMIT	1.00	0.23	1.00
98	ALERT_INACTIVITY_EXIT	0.08	0.09	0.08
99	NAVIGATION_PROFILE_LOGIN	0.37	0.33	0.41
	End Token	0.35	0.03	0.28
	MAI	**0.64**	**0.25**	**0.60**

Table 11 Example of clickstream - Repeated actions of "ITEM BOOKMARK ON" and " ITEM BOOKMARK OFF"

		Predicted Probability		
	The Clickstream Sequence	LSTM	MCNA	RNN
1	NAVIGATION_PROFILE_LOGIN	0.94	0.93	0.93
2	NAVIGATION_PROFILE_CHOOSE	0.90	0.78	0.93

3	NAVIGATION_ACCESS_CODE_SUBMIT	0.93	0.90	0.90
4	NAVIGATION_DIRECTIONS_CONTINUE	0.84	0.75	0.90
5	NAVIGATION_ITEM_NEXT	0.22	0.25	0.19
6	NAVIGATION_ITEM_BACK	0.94	0.07	0.88
7	ITEM_BOOKMARK_ON	0.67	0.11	0.66
8	NAVIGATION_REVIEW_PANEL_OPEN	0.77	0.41	0.77
9	NAVIGATION_REVIEW_PANEL_CLOSE	0.95	0.75	0.94
10	NAVIGATION_ITEM_JUMP	0.73	0.46	0.85
11	NAVIGATION_REVIEW_PANEL_OPEN	0.11	0.25	0.13
12	NAVIGATION_REVIEW_PANEL_CLOSE	0.98	0.75	0.98
13	NAVIGATION_ITEM_JUMP	0.85	0.46	0.88
14	NAVIGATION_REVIEW_PANEL_OPEN	0.17	0.25	0.23
15	NAVIGATION_REVIEW_PANEL_CLOSE	0.98	0.75	0.99
16	NAVIGATION_ITEM_JUMP	0.87	0.46	0.90
17	NAVIGATION_REVIEW_PANEL_OPEN	0.13	0.25	0.33
18	NAVIGATION_REVIEW_PANEL_CLOSE	0.98	0.75	0.99
19	NAVIGATION_ITEM_JUMP	0.86	0.46	0.92
20	**ITEM_BOOKMARK_OFF**	**0.82**	**0.19**	**0.28**
21	**ITEM_BOOKMARK_ON**	**0.05**	**0.20**	**0.09**
22	**ITEM_BOOKMARK_OFF**	**0.85**	**0.32**	**0.80**
23	**ITEM_BOOKMARK_ON**	**0.35**	**0.20**	**0.42**
24	**ITEM_BOOKMARK_OFF**	**0.94**	**0.32**	**0.89**
25	**ITEM_BOOKMARK_ON**	**0.72**	**0.20**	**0.78**
26	**ITEM_BOOKMARK_OFF**	**0.96**	**0.32**	**0.92**
27	**ITEM_BOOKMARK_ON**	**0.84**	**0.20**	**0.85**
28	**ITEM_BOOKMARK_OFF**	**0.97**	**0.32**	**0.93**
29	**ITEM_BOOKMARK_ON**	**0.87**	**0.20**	**0.87**
30	**ITEM_BOOKMARK_OFF**	**0.97**	**0.32**	**0.93**
31	**ITEM_BOOKMARK_ON**	**0.89**	**0.20**	**0.87**
32	**ITEM_BOOKMARK_OFF**	**0.97**	**0.32**	**0.93**
33	**ITEM_BOOKMARK_ON**	**0.90**	**0.20**	**0.87**
34	**ITEM_BOOKMARK_OFF**	**0.97**	**0.32**	**0.93**
35	**ITEM_BOOKMARK_ON**	**0.90**	**0.20**	**0.87**
36	**ITEM_BOOKMARK_OFF**	**0.97**	**0.32**	**0.93**
37	**ITEM_BOOKMARK_ON**	**0.90**	**0.20**	**0.87**
38	**ITEM_BOOKMARK_OFF**	**0.97**	**0.32**	**0.93**
39	**ITEM_BOOKMARK_ON**	**0.90**	**0.20**	**0.87**
40	**ITEM_BOOKMARK_OFF**	**0.97**	**0.32**	**0.93**
41	**ITEM_BOOKMARK_ON**	**0.91**	**0.20**	**0.87**
42	**ITEM_BOOKMARK_OFF**	**0.97**	**0.32**	**0.93**
43	**ITEM_BOOKMARK_ON**	**0.91**	**0.20**	**0.87**
44	**ITEM_BOOKMARK_OFF**	**0.98**	**0.32**	**0.93**
45	**ITEM_BOOKMARK_ON**	**0.91**	**0.20**	**0.87**
46	**ITEM_BOOKMARK_OFF**	**0.98**	**0.32**	**0.93**
47	**ITEM_BOOKMARK_ON**	**0.91**	**0.20**	**0.87**
48	**ITEM_BOOKMARK_OFF**	**0.98**	**0.32**	**0.93**

49	**ITEM_BOOKMARK_ON**	**0.91**	**0.20**	**0.87**
50	**ITEM_BOOKMARK_OFF**	**0.97**	**0.32**	**0.93**
51	**ITEM_BOOKMARK_ON**	**0.91**	**0.20**	**0.87**
52	**ITEM_BOOKMARK_OFF**	**0.97**	**0.32**	**0.93**
53	**ITEM_BOOKMARK_ON**	**0.91**	**0.20**	**0.87**
54	**ITEM_BOOKMARK_OFF**	**0.97**	**0.32**	**0.93**
55	**ITEM_BOOKMARK_ON**	**0.91**	**0.20**	**0.87**
56	**ITEM_BOOKMARK_OFF**	**0.97**	**0.32**	**0.93**
57	**ITEM_BOOKMARK_ON**	**0.91**	**0.20**	**0.87**
58	**ITEM_BOOKMARK_OFF**	**0.97**	**0.32**	**0.93**
59	**ITEM_BOOKMARK_ON**	**0.91**	**0.20**	**0.87**
60	**ITEM_BOOKMARK_OFF**	**0.97**	**0.32**	**0.93**
61	**ITEM_BOOKMARK_ON**	**0.90**	**0.20**	**0.87**
62	**ITEM_BOOKMARK_OFF**	**0.97**	**0.32**	**0.93**
63	**ITEM_BOOKMARK_ON**	**0.90**	**0.20**	**0.87**
64	**ITEM_BOOKMARK_OFF**	**0.97**	**0.32**	**0.93**
65	**ITEM_BOOKMARK_ON**	**0.90**	**0.20**	**0.87**
66	**ITEM_BOOKMARK_OFF**	**0.97**	**0.32**	**0.93**
67	**ITEM_BOOKMARK_ON**	**0.90**	**0.20**	**0.87**
68	**ITEM_BOOKMARK_OFF**	**0.97**	**0.32**	**0.93**
69	**ITEM_BOOKMARK_ON**	**0.90**	**0.20**	**0.87**
70	**ITEM_BOOKMARK_OFF**	**0.97**	**0.32**	**0.93**
71	**ITEM_BOOKMARK_ON**	**0.90**	**0.20**	**0.87**
72	**ITEM_BOOKMARK_OFF**	**0.97**	**0.32**	**0.93**
73	**ITEM_BOOKMARK_ON**	**0.89**	**0.20**	**0.87**
74	**ITEM_BOOKMARK_OFF**	**0.97**	**0.32**	**0.93**
75	**ITEM_BOOKMARK_ON**	**0.89**	**0.20**	**0.87**
76	**ITEM_BOOKMARK_OFF**	**0.97**	**0.32**	**0.93**
77	**ITEM_BOOKMARK_ON**	**0.89**	**0.20**	**0.87**
78	**ITEM_BOOKMARK_OFF**	**0.97**	**0.32**	**0.93**
79	**ITEM_BOOKMARK_ON**	**0.89**	**0.20**	**0.87**
80	**ITEM_BOOKMARK_OFF**	**0.97**	**0.32**	**0.93**
81	TOOL_SKETCH_CLOSE	0.01	0.00	0.00
82	TOOL_TEXT_HIGHLIGHT_TOGGLE	0.10	0.37	0.50
83	TOOL_TEXT_HIGHLIGHT_SELECTED	0.50	0.26	0.20
84	TOOL_TEXT_HIGHLIGHT_CANCEL_ALL	0.36	0.29	0.33
85	TOOL_TEXT_HIGHLIGHT_TOGGLE	0.79	0.59	0.70
86	TOOL_CALCULATOR_TOGGLE	0.42	0.19	0.54
87	TOOL_CALCULATOR_OPEN	0.53	0.57	0.48
88	TOOL_CALCULATOR_CLOSE	0.91	0.49	0.91
89	TOOL_REFERENCES_TOGGLE	0.43	0.07	0.61
90	TOOL_REFERENCES_OPEN	0.83	0.62	0.79
91	TOOL_REFERENCES_CLOSE	0.92	0.67	0.87
92	NAVIGATION_ITEM_NEXT	0.25	0.16	0.24
93	ITEM_MULTIPLE_CHOICE_ANSWER	0.61	0.57	0.62
94	NAVIGATION_ITEM_NEXT	0.78	0.70	0.77

95	ITEM_MULTIPLE_CHOICE_ANSWER	0.47	0.57	0.05
96	NAVIGATION_ITEM_NEXT	0.77	0.70	0.80
97	ITEM_MULTIPLE_CHOICE_ANSWER	0.75	0.57	0.40
98	NAVIGATION_ITEM_NEXT	0.81	0.70	0.84
99	ITEM_MULTIPLE_CHOICE_ANSWER	0.76	0.57	0.69
100	NAVIGATION_ITEM_NEXT	0.83	0.70	0.84
101	ITEM_MULTIPLE_CHOICE_ANSWER	0.76	0.57	0.75
102	NAVIGATION_ITEM_NEXT	0.82	0.70	0.80
103	ITEM_MULTIPLE_CHOICE_ANSWER	0.75	0.57	0.76
104	NAVIGATION_ITEM_NEXT	0.82	0.70	0.79
105	NAVIGATION_REVIEW_PANEL_OPEN	0.02	0.04	0.05
106	NAVIGATION_REVIEW_PANEL_CLOSE	0.84	0.75	0.98
107	ITEM_TILE_BOX_DRAG_START	0.02	0.00	0.00
108	ITEM_TILE_BOX_DRAG_END	0.98	1.00	0.97
109	ITEM_TILE_BOX_DRAG_START	0.92	0.79	0.91
110	ITEM_TILE_BOX_DRAG_END	1.00	1.00	1.00
111	ITEM_TILE_BOX_DRAG_START	0.87	0.79	0.94
112	ITEM_TILE_BOX_DRAG_END	0.99	1.00	1.00
113	NAVIGATION_ITEM_NEXT	0.28	0.16	0.25
114	ITEM_MULTIPLE_CHOICE_ANSWER	0.82	0.57	0.79
115	NAVIGATION_ITEM_NEXT	0.69	0.70	0.80
116	ITEM_MULTIPLE_CHOICE_ANSWER	0.77	0.57	0.77
117	NAVIGATION_ITEM_NEXT	0.70	0.70	0.86
118	ITEM_MULTIPLE_CHOICE_ANSWER	0.81	0.57	0.85
119	NAVIGATION_REVIEW_PANEL_OPEN	0.27	0.04	0.01
120	NAVIGATION_TURN_IN_START	0.88	0.23	0.62
121	NAVIGATION_REVIEW_PANEL_CLOSE	0.97	0.98	0.99
122	NAVIGATION_TURN_IN_COMMIT	1.00	0.23	1.00
	End Token	0.34	0.21	0.44
	MAI	**0.79**	**0.39**	**0.77**

Discussion

This study evaluated the performance of three behavior sequence prediction models: LSTM, RNN, and MCNA (bigram). The MAI statistic was defined and used to quantify 'typical' and 'atypical' test-taking behaviors in clickstreams. Among the three models, the LSTM model had the highest prediction accuracy compared to the two baseline approaches. MCNA and LSTM sometimes generated different MAI results, especially when repeated actions occur during testing.

The MAI indices are also compared to students' performance and other traditional aberrance detection indicators. Results show that students with the lowest and highest achievements show more typical behavior patterns, while students in the middle level of performance have more atypical behaviors. However, the amount of MAI difference is relatively small across the performance groups. This finding is to some extent expected. Unlike the process data from problem-solving items, the clickstream actions for multiple-choice items are less likely to be related to students' performance. On the other hand, MAI is moderately negatively correlated with answer change indices. When an examinee changes the answers for many times, MAI will identify the clickstream as atypical. The MAI based on LSTM is more correlated with these indices, compared to the MAI based on MCNA.

In addition, atypical behavior patterns are identified in the clickstreams with low MAI scores. In our case study analysis of a low MAI clickstream, the test-taker apparently repeatedly opened and closed each of the available tools on the first item before answering it. Such behavior is very uncommon among all the test-takers. Moreover, we compared the action frequencies between low MAI and high MAI groups. The most common "typical" and "atypical" actions and their frequency were substantially different between low and high MAI groups. Quite a few mismatching predictions were related to tool usage. For example, calculator toggle was observed more commonly in the low "MAI" group, appearing more rarely in the high MAI group.

This study is limited in several ways. Firstly, the clickstream data in this study comes from only one test session in a math summative assessment. The test consists of multiple-choice items and technology-enhanced items only. Thus, the findings from this study might not generalize to different tests. Secondly, it is possible that the data of some clickstreams was corrupted and is missing data in unpredictable ways. Clickstream data are typically collected from a test delivery system where tens of thousands of clickstreams might be tracked at the same time. In the current data file, we noticed missing information on some students' login actions. However, missingness in other parts of the clickstream is more difficult to detect. To decrease the impact of data missingness, we removed clickstreams with extremely short length (less than 30 actions) in this study. Finally, interpreting the behavioral predictive model results are less straightforward compared to models where input features are more strictly defined. The LSTM model does not explain why one individual's clickstream achieves a high MAI and a different one achieves a low MAI. Since the model depends entirely on the training data and the distribution of behaviors in the training data, the interpretations about what "low" or "high" MAI means in terms of actual behaviors will always depend on post-hoc analysis of examinee behavior clickstreams at varying levels of MAI. In all circumstances, a low MAI indicates that the behaviors of an individual were less expected relative to the population of other test-takers.

The overarching goal of this line of research is to be able to quantify how "typical" or "atypical" a test-takers' behaviors are. When something "atypical" happens, then stakeholders can identify what is going on and determine whether any remediation or action is necessary. In the current study, an LSTM approach towards behavior modeling was proposed, borrowing from sequence prediction methods that have been utilized in the rapidly advancing language modeling field. LSTM approaches allow for prediction models to learn exclusively from the training data, rather than relying on any engineered, pre-conceived notion of what behavior patterns ought to be. A downstream application of the proposed methodology would be to apply it as an additional surveying or monitoring technique, in conjunction with other process data and test security analysis protocols. Future studies could improve upon the current study by collecting more precise clickstream data, including response time information in the behavior prediction models, or using alternative sequence behavior prediction models. It would also be an interesting study to apply MAI to other types of clickstream data, including more complex process data from interactive problem-solving items or collaborative tasks.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., . . . Jia, Y. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Retrieved from tensorflow.org

Banerjee, A., & Ghosh, J. (2011). Clickstream Clustering Using Weighted Longest Common Subsequences. *Proc. of the Web Mining Workshop in CDM.* Retrieved from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.29.427&rep=rep1&type=pdf

Bishop, S., & Egan, K. (2017). Detecting erasures and unusual gain scores: Understanding the status quo. In *G. J. Cizek & J. A. Wollack (Eds.). Handbook of Quantitative Methods for Detecting Cheating on Tests* (pp. 193-213). Washington, DC: Routledge. Retrieved from https://www.taylorfrancis.com/chapters/edit/10.4324/9781315743097-10/detecting-erasures-unusual-gain-scores-scott-bishop-karla-egan

Chollet, F., & Others. (2015). Keras. Retrieved from https://github.com/keras-team/keras

Drasgrow, F., Levine, V. M., & Williams, A. E. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67-86. Retrieved from https://psycnet.apa.org/record/1985-24320-001

Graves, A. (2014). Generating Sequences with Recurrent Neural Networks. *ArXiv*, 1-43. doi:https://doi.org/10.48550/arXiv.1308.0850

Gunduz, S., & Ozsu, M. T. (2003). A Web page prediction model based on click-stream tree representation of user behavior. *Proc. of KDD.* Retrieved from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.3.1067&rep=rep1&type=pdf

He, Q., Liao, D., & Jiao, H. (2019). Clustering behavioral patterns using procee data in PIAAC problem-solving items. In B. Veldkamp, & C. Sluijter, *Theoretical and practical advances in computer-based educational measurement.* (pp. 189-212). Cham: Springer. doi:https://doi.org/10.1007/978-3-030-18480-3_10

Heer, J., & Chi, E. H. (2002). Mining the Structure of User Activity using Cluster Stability. *Proc. of the Workshop on Web Analytics, SIAM Conference on Data Mining.* Retrieved from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.16.3665&rep=rep1&type=pdf

Li, Z., Wall, N., & Tang, H. (2018). A new statistic for detecting aberrant response time patterns in large-scale assessments. *Paper presented at the annual meeting of the National Council on Measurement in Education (NCME).* New York, NY. Retrieved from https://www.emetric.net/Content/pdf/Manuscript-Response%20Aberrance.pdf

Liao, M., Patton, J., Yan, R., & Jiao, H. (2021). Mining process data to detect aberrant test takers. *Measurement: Interdisciplinary research and perspectives, 19*(2), 93-105.

Ranger, J., Schmidt, N., & Wolgast, A. (2020). The detection of cheating on e-Exams in higher education-The performance of several old and some new indicators. *frontiers in Psychology, 11*. Retrieved from https://doi.org/10.3389/fpsyg.2020.568825

Su, Q., & Chen, L. (2015). A Method for Discovering Clusters of E-commerce Interest Patterns using Click-stream Data. *ECRA*, *14*, pp. 1-13. Retrieved from https://doi.org/10.1016/j.elerap.2014.10.002

Sundermeyer, M., Schlüter, R., & Ney, H. (2012). LSTM Neural Networks for Language Modeling. *Interspeech.* Retrieved from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.248.4448&rep=rep1&type=pdf

Tang, S., Peterson, J., & Pardos, Z. (2017). Predictive Modeling of Student Behavior Using Granular Large Scale Action Data from a MOOC. Retrieved from https://www.researchgate.net/publication/315874906_Predictive_Modelling_of_Student_Behaviour_Using_Granular_Large-Scale_Action_Data

Tang, X., Wang, Z., He, Q., Liu, J., & Ying, Z. (2020). Latent feature extraction for process data via multidimensional scaling. *Psychometrika, 85*, 378-397.

Tang, X., Wang, Z., Liu, J., & Ying, Z. (2020). An exploratory analysis of the latent structure of process data via action sequence autoencoder. *British Journal of Mathematical and Statistical Psychology, 74*(1), 1-33. doi: https://doi.org/10.1111/bmsp.12203

van der Linden, J. W., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika, 73*(3), 365-384. Retrieved from https://doi.org/10.1007/s11336-007-9046-8

Wang, G., Zhang, X., Tang, S., Wilson, C., Zheng, H., & Zhao, B. Y. (2017). *Clickstream User Behavior Models.* ACM Transactions on the Web. Retrieved from https://doi.org/10.1145/3068332

Wise, S., & DeMars, C. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, 19-38. Retrieved from https://www.jstor.org/stable/20461807

Appendix

Table 12 Full List of Mismatched observed and predicted clickstream actions of the "Low MAI" group

Observed Action	Predicted Action by LSTM	N	Percent
ITEM_MULTIPLE_CHOICE_ANSWER	NAVIGATION_ITEM_NEXT	397	4.5%
TOOL_CALCULATOR_TOGGLE	ITEM_MULTIPLE_CHOICE_ANSWER	227	2.6%
NAVIGATION_ITEM_NEXT	ITEM_MULTIPLE_CHOICE_ANSWER	223	2.5%
TOOL_ANSWER_MASKING_TOGGLE	ITEM_MULTIPLE_CHOICE_ANSWER	168	1.9%
TOOL_CALCULATOR_CLOSE	ITEM_MULTIPLE_CHOICE_ANSWER	144	1.6%
TOOL_CALCULATOR_TOGGLE	TOOL_CALCULATOR_OPEN	134	1.5%
TOOL_ANSWER_MASKING_TOGGLE	NAVIGATION_ITEM_NEXT	115	1.3%
NAVIGATION_REVIEW_PANEL_OPEN	NAVIGATION_ITEM_NEXT	104	1.2%
NAVIGATION_REVIEW_PANEL_OPEN	ITEM_MULTIPLE_CHOICE_ANSWER	99	1.1%
NAVIGATION_ITEM_BACK	ITEM_MULTIPLE_CHOICE_ANSWER	96	1.1%
ITEM_MULTIPLE_CHOICE_ANSWER	TOOL_CALCULATOR_CLOSE	90	1.0%
NAVIGATION_ITEM_BACK	NAVIGATION_ITEM_NEXT	79	0.9%
ITEM_MULTIPLE_CHOICE_ANSWER	TOOL_ANSWER_MASKING_TOGGLE	78	0.9%
NAVIGATION_ITEM_NEXT	NAVIGATION_ITEM_BACK	73	0.8%
TOOL_CALCULATOR_CLOSE	NAVIGATION_ITEM_NEXT	66	0.7%
NAVIGATION_ITEM_NEXT	TOOL_ANSWER_MASKING_TOGGLE	58	0.7%
TOOL_SKETCH_CLOSE	TOOL_SKETCH_SELECT	54	0.6%
TOOL_CALCULATOR_TOGGLE	NAVIGATION_ITEM_NEXT	51	0.6%
TOOL_REFERENCES_TOGGLE	ITEM_MULTIPLE_CHOICE_ANSWER	50	0.6%
ITEM_MULTIPLE_CHOICE_ANSWER	NAVIGATION_ITEM_BACK	44	0.5%
TOOL_REFERENCES_TOGGLE	TOOL_REFERENCES_OPEN	41	0.5%
ITEM_MULTIPLE_CHOICE_ANSWER	NAVIGATION_REVIEW_PANEL_OPEN	38	0.4%
TOOL_CALCULATOR_OPEN	TOOL_CALCULATOR_TOGGLE	38	0.4%
ITEM_SELECT_DROP_DOWN_select	ITEM_MULTIPLE_CHOICE_ANSWER	34	0.4%
NAVIGATION_ITEM_NEXT	NAVIGATION_REVIEW_PANEL_OPEN	34	0.4%
NAVIGATION_REVIEW_PANEL_CLOSE	NAVIGATION_TURN_IN_START	33	0.4%
NAVIGATION_ITEM_NEXT	TOOL_CALCULATOR_CLOSE	29	0.3%
NAVIGATION_TURN_IN_START	NAVIGATION_REVIEW_PANEL_CLOSE	27	0.3%
TOOL_CALCULATOR_TOGGLE	TOOL_CALCULATOR_CLOSE	25	0.3%
NAVIGATION_ITEM_NEXT	ITEM_TILE_BOX_DRAG_START	24	0.3%
NAVIGATION_PROFILE_CHOOSE	NAVIGATION_PROFILE_LOGIN	23	0.3%
End Token	ALERT_PROFILE_EXIT	23	0.3%
ITEM_MULTIPLE_CHOICE_ANSWER	TOOL_CALCULATOR_TOGGLE	22	0.2%
NAVIGATION_ITEM_NEXT	ITEM_SELECT_DROP_DOWN_select	22	0.2%
ITEM_BOOKMARK_OFF	NAVIGATION_ITEM_NEXT	21	0.2%
ITEM_BOOKMARK_ON	ITEM_MULTIPLE_CHOICE_ANSWER	20	0.2%
TOOL_CALCULATOR_TOGGLE	NAVIGATION_ITEM_BACK	20	0.2%
NAVIGATION_REVIEW_PANEL_OPEN	NAVIGATION_ITEM_JUMP	18	0.2%
TOOL_SKETCH_OPEN	ITEM_MULTIPLE_CHOICE_ANSWER	18	0.2%
ITEM_TILE_BOX_DRAG_START	ITEM_MULTIPLE_CHOICE_ANSWER	17	0.2%
NAVIGATION_PROFILE_LOGIN	End Token	17	0.2%
ITEM_BOOKMARK_ON	NAVIGATION_ITEM_NEXT	16	0.2%

ITEM_MULTIPLE_CHOICE_ANSWER	TOOL_REFERENCES_CLOSE	15	0.2%
TOOL_ANSWER_MASKING_TOGGLE	NAVIGATION_ITEM_BACK	15	0.2%
TOOL_CALCULATOR_TOGGLE	TOOL_ANSWER_MASKING_TOGGLE	15	0.2%
ALERT_INACTIVITY_EXIT	ALERT_PROFILE_EXIT	14	0.2%
NAVIGATION_ITEM_JUMP	ITEM_MULTIPLE_CHOICE_ANSWER	14	0.2%
TOOL_REFERENCES_OPEN	TOOL_REFERENCES_TOGGLE	14	0.2%
ITEM_MULTIPLE_CHOICE_ANSWER	NAVIGATION_ITEM_JUMP	13	0.1%
TOOL_CALCULATOR_CLOSE	NAVIGATION_ITEM_BACK	13	0.1%
TOOL_REFERENCES_CLOSE	ITEM_MULTIPLE_CHOICE_ANSWER	13	0.1%
TOOL_REFERENCES_TOGGLE	NAVIGATION_ITEM_NEXT	13	0.1%
ITEM_MULTIPLE_CHOICE_ANSWER	ITEM_SELECT_DROP_DOWN_select	12	0.1%
ITEM_SELECT_DROP_DOWN_select	NAVIGATION_ITEM_NEXT	12	0.1%
TOOL_TEXT_HIGHLIGHT_TOGGLE	ITEM_MULTIPLE_CHOICE_ANSWER	12	0.1%
NAVIGATION_ITEM_NEXT	NAVIGATION_ITEM_JUMP	11	0.1%
TOOL_REFERENCES_TOGGLE	TOOL_CALCULATOR_OPEN	11	0.1%
TOOL_REFERENCES_TOGGLE	TOOL_CALCULATOR_TOGGLE	11	0.1%
ALERT_INACTIVITY_EXIT	End Token	10	0.1%
ALERT_PROFILE_EXIT	End Token	10	0.1%
ITEM_BOOKMARK_OFF	NAVIGATION_REVIEW_PANEL_OPEN	10	0.1%
ITEM_MULTIPLE_CHOICE_ANSWER	ITEM_DRAG_BOX_DRAG_START	10	0.1%
ITEM_MULTIPLE_CHOICE_ANSWER	TOOL_CALCULATOR_OPEN	10	0.1%
NAVIGATION_ACCESS_CODE_SUBMIT	NAVIGATION_PROFILE_CHOOSE	10	0.1%
NAVIGATION_DIRECTIONS_CONTINUE	NAVIGATION_PROFILE_CHOOSE	10	0.1%
NAVIGATION_REVIEW_PANEL_OPEN	TOOL_CALCULATOR_CLOSE	10	0.1%

- Note: The events with less than 10 counts are removed from the list.

Table 13 Clickstream Action List

Action	Code of Action
NULL_RECORD	0
ALERT_DIRECTIONS_EXIT	1
ALERT_DIRE_WARNING_CLOSE	2
ALERT_DIRE_WARNING_RETRY	3
ALERT_FINAL_SCORE_UNAVAILABLE_CLOSE	4
ALERT_INACTIVITY_EXIT	5
ALERT_LOCK_TIMEOUT_EXIT	6
ALERT_OFFLINE_WARNING_CLOSE	7
ALERT_OFFLINE_WARNING_READ	8
ALERT_PROCTOR_PASSWORD_SUBMIT	9
ALERT_PROFILE_EXIT	10
ALERT_SIMULTANEOUS_USER_CLOSE	11
ALERT_START_TEST_ERROR_CLOSE	12
ALERT_START_TEST_ERROR_RETRY	13
ALERT_TIMEOUT_CLOSE	14
ALERT_TTS_FAILURE_CLOSE	15
ITEM_BOOKMARK_OFF	16

ITEM_BOOKMARK_ON	17
ITEM_CLEAR_CANCEL	18
ITEM_CLEAR_COMMIT	19
ITEM_CLEAR_START	20
ITEM_CONNECTION_match	21
ITEM_CONNECTION_unmatch	22
ITEM_DRAG_BOX_DRAG_END	23
ITEM_DRAG_BOX_DRAG_START	24
ITEM_HOTSPOT_select	25
ITEM_HOTSPOT_unselect	26
ITEM_MATH_EQUATION_CANCEL	27
ITEM_MATH_EQUATION_OPEN	28
ITEM_MATH_EQUATION_SELECT	29
ITEM_MULTIPLE_CHOICE_ANSWER	30
ITEM_MULTIPLE_CHOICE_Eliminate	31
ITEM_MULTIPLE_CHOICE_UnEliminate	32
ITEM_OPEN_ENDED_BLUR	33
ITEM_OPEN_ENDED_BOLD	34
ITEM_OPEN_ENDED_COPY	35
ITEM_OPEN_ENDED_CUT	36
ITEM_OPEN_ENDED_FOCUS	37
ITEM_OPEN_ENDED_ITALIC	38
ITEM_OPEN_ENDED_PASTE	39
ITEM_OPEN_ENDED_REDO	40
ITEM_OPEN_ENDED_SPELLCHECK_OFF	41
ITEM_OPEN_ENDED_SPELLCHECK_ON	42
ITEM_OPEN_ENDED_UNDERLINE	43
ITEM_OPEN_ENDED_UNDO	44
ITEM_SELECTTEXT_select	45
ITEM_SELECTTEXT_unselect	46
ITEM_SELECT_DROP_DOWN_select	47
ITEM_STIMULUS_SELECT	48
ITEM_STIMULUS_TOGGLE	49
ITEM_TILE_BOX_DRAG_END	50
ITEM_TILE_BOX_DRAG_START	51
NAVIGATION_ACCESS_CODE_CANCEL	52
NAVIGATION_ACCESS_CODE_SUBMIT	53
NAVIGATION_ACCOMMODATION_OPTIONS_CONTINUE	54
NAVIGATION_DIRECTIONS_ACCOMMODATION_CLOSE	55
NAVIGATION_DIRECTIONS_ACCOMMODATION_OPEN	56
NAVIGATION_DIRECTIONS_CONTINUE	57
NAVIGATION_FINAL_SCORE_CLOSE	58
NAVIGATION_ITEM_BACK	59
NAVIGATION_ITEM_JUMP	60
NAVIGATION_ITEM_NEXT	61

NAVIGATION_LOCK_RESUME	62
NAVIGATION_LOCK_SIGN_OUT	63
NAVIGATION_PAUSE_CANCEL	64
NAVIGATION_PAUSE_COMMIT	65
NAVIGATION_PAUSE_LOCK	66
NAVIGATION_PROFILE_CHOOSE	67
NAVIGATION_PROFILE_LOGIN	68
NAVIGATION_REVIEW_PANEL_CLOSE	69
NAVIGATION_REVIEW_PANEL_OPEN	70
NAVIGATION_SECTION_DENIED_CLOSE	71
NAVIGATION_SECTION_WARNING_CANCEL	72
NAVIGATION_SECTION_WARNING_COMMIT	73
NAVIGATION_SHOW_ANSWER_CLOSE	74
NAVIGATION_SHOW_ANSWER_OPEN	75
NAVIGATION_SHOW_ANSWER_SELECT	76
NAVIGATION_TURN_IN_CANCEL	77
NAVIGATION_TURN_IN_COMMIT	78
NAVIGATION_TURN_IN_START	79
NAVIGATION_trigger_START	80
TOOL_ANSWER_MASKING_DISABLE	81
TOOL_ANSWER_MASKING_ENABLE	82
TOOL_ANSWER_MASKING_TOGGLE	83
TOOL_CALCULATOR_CLOSE	84
TOOL_CALCULATOR_OPEN	85
TOOL_CALCULATOR_TOGGLE	86
TOOL_COLOR_SCHEME_DISABLE	87
TOOL_COLOR_SCHEME_ENABLE	88
TOOL_COLOR_SCHEME_OFF	89
TOOL_COLOR_SCHEME_ON	90
TOOL_COLOR_SCHEME_TOGGLE	91
TOOL_CUSTOM_MASKING_CLOSE	92
TOOL_CUSTOM_MASKING_DISABLE	93
TOOL_CUSTOM_MASKING_ENABLE	94
TOOL_CUSTOM_MASKING_OPEN	95
TOOL_CUSTOM_MASKING_TOGGLE	96
TOOL_DICTIONARY_CLOSE	97
TOOL_DICTIONARY_OPEN	98
TOOL_DICTIONARY_TOGGLE	99
TOOL_Eliminator_DISABLE	100
TOOL_Eliminator_ENABLE	101
TOOL_GUIDELINE_CLOSE	102
TOOL_GUIDELINE_DISABLE	103
TOOL_GUIDELINE_ENABLE	104
TOOL_GUIDELINE_OPEN	105
TOOL_MASKING_DISABLE	106
TOOL_MASKING_ENABLE	107
TOOL_NOTEPAD_BLUR	108
TOOL_NOTEPAD_CLOSE	109

TOOL_NOTEPAD_OPEN	110
TOOL_PROTRACTOR_CLOSE	111
TOOL_PROTRACTOR_OPEN	112
TOOL_REFERENCES_CLOSE	113
TOOL_REFERENCES_OPEN	114
TOOL_REFERENCES_TOGGLE	115
TOOL_REVERSE_CONTRAST_DISABLE	116
TOOL_REVERSE_CONTRAST_ENABLE	117
TOOL_REVERSE_CONTRAST_OFF	118
TOOL_REVERSE_CONTRAST_ON	119
TOOL_RULER_CLOSE	120
TOOL_RULER_OPEN	121
TOOL_RULER_TOGGLE	122
TOOL_SIGNING_DISABLE	123
TOOL_SIGNING_ENABLE	124
TOOL_SKETCH_CLOSE	125
TOOL_SKETCH_OPEN	126
TOOL_SKETCH_SELECT	127
TOOL_TEXT_HIGHLIGHT_CANCEL	128
TOOL_TEXT_HIGHLIGHT_CANCEL_ALL	129
TOOL_TEXT_HIGHLIGHT_SELECTED	130
TOOL_TEXT_HIGHLIGHT_TOGGLE	131
TOOL_TTS_DISABLE	132
TOOL_TTS_ENABLE	133
TOOL_TTS_OFF	134
TOOL_TTS_ON	135
TOOL_TTS_RATE	136
TOOL_TTS_SELECT	137
TOOL_TTS_VOLUME	138
TOOL_ZOOM_DECREASE	139
TOOL_ZOOM_DISABLE	140
TOOL_ZOOM_ENABLE	141
TOOL_ZOOM_INCREASE	142
TOOL_ZOOM_RESET	143
TOOL_ZOOM_SET	144
NAVIGATION_REVIEW_PANEL_START	145
NAVIGATION_TOOLBAR_START	146
TOOL_TTS_PAUSE	147
TOOL_TTS_PLAY	148
TOOL_TTS_RESUME	149
TOOL_TTS_SKIP	150
TOOL_TTS_STOP	151