Evaluating Effectiveness of Gang Reduction Youth Development Program With Dynamic Mode Decomposition and Machine Learning Techniques

Zehan Chao[a], Zheyuan Cui[a], Avery Edson[a], Cesar Guajardo[b], Yihuan Huang[a], Xingjia Wang[a], Zhanyuan Yin[a], Heather Z. Brooks[a], P. Jeffrey Brantingham[a], and Andrea Bertozzi[a]

[a]University of California, Los Angeles; [b]University of California, San Diego

This manuscript was compiled on August 16, 2019

The Gang Reduction and Youth Development (GRYD) program is an initiative conducted by the mayor's office of the City of Los Angeles, with the aim to curb gang violence and to promote youth development among at-risk individuals. Eligibility for program services is established using the Youth Services Eligibility Tool (YSET) questionnaire. The goal of the study is to evaluate the effectiveness of the GRYD program through a dataset that records the participants' responses to the YSET questionnaire. Existing machine learning algorithms, such as Linear Support Vector Machine (LSVM) and Neural Network (NN), can help us accurately predict future responses and risk-indicating scores of the GRYD program participants, yet interpreting the resulting models of these algorithms remains difficult. On the other hand, through the use of dynamical models such as Dynamic Mode Decomposition (DMD) and Dynamic Mode Decomposition with Control (DMDc), we were able to not only predict individual responses and risk scores as accurately as by using machine learning models, but also interpret how responses to each question change over time in different demographic groups of participants. In addition, we were able to observe how significant each question is when determining participants' overall risk scores. To further improve the effectiveness of program, we suggest GRYD to consider targeting services in different risk-influencing areas differently for participants based on program progress and their distinct demographic groups.

dynamic mode decomposition | machine learning | data analysis | gang reduction

Introduction

The Gang Reduction and Youth Development Program (GRYD) is a prevention program run by the mayor's office of the City of Los Angeles that dedicates their efforts to curb gang violence and promote youth development among at-risk children and young adolescents. This program is divided into 23 different districts located throughout the City of Los Angeles, primarily focused in high-risk neighborhoods. To determine eligibility, GRYD administers two different sets of questionnaires, based on the participant's age. One of them is the Youth Services Eligibility Tool (YSET) which is geared for children ages 10 to 16 years old. For the purposes of this project, we worked exclusively on YSET due to a higher sample size to conduct our model (1).

YSET Questionnaire. Eligibility for this program is determined by the Youth Services Eligibility Tool (YSET) questionnaire, which is aimed for children ages 10-16 years old. This questionnaire consists of a total of 104 questions, among which 56 are eligibility-determining. These questions can be separated into 17 delinquency-based behavioral questions and 39 opinion-based attitudinal questions. All questions can be grouped into 9 sections, each measuring a factor of the participants' risk level (e.g. parental supervision, impulsive risk taking, peer delinquency). Scores from each factor are summed up and mapped into sectional concern levels (0-4) according to a set of scoring instructions, and then combined to calculate a comprehensive risk score (0-9, one point for each section of which the corresponding concern level is greater than 0). Participants who obtain risk scores of 4 or higher are eligible for the GRYD secondary (full) prevention program, and those with risk scores lower than 4 are eligible for a primary (partial) prevention program. All participants referred to the GRYD program are required to take this questionnaire to determine if they are eligible for a full program, and this response is labelled as their "intake" response; participants who have been enrolled in the secondary prevention program are expected to take the questionnaire every six months while they stay in the program, and these responses are labelled as their "retake" responses.

	Sect.	Risk-Influencing Factor	Questions	Sample Question from Section
Attitudinal	A	Antisocial Tendencies	A1 - A6	A1: "I try to be nice to other people because I care about their feelings."
	B	Weak Parental Supervision	B7 - B9	B7: "When I am not at home or school, my parents or guardians know where I am."
	C	Critical Life Events	C10 - C16	C14: "...did you have a big fight or problem with a friend?"
	DE	Impulsive Risk Taking	DE17 - DE20	DE17: "Sometimes I like to do something dangerous just for the fun of it."
	F	Neutralization	F21 - F26	F25: "It is okay to beat people up if they hit me first."
	G	Negative Peer Influence	G27 - G31	G29: "If your friends were getting you into trouble at home, would you still hang out with them?"
	H	Peer Delinquency	H32 - H37	H33: "How many of your friends have stolen something?"
	T	Family Gang Influence	T38 - T39	T39: "How many people in your family are gang members?"
Behavioral	IJ	Self-Reported Delinquency & Substance Use	IJ40 - IJ56	IJ47: "Have you carried a hidden weapon for protection ever/in 6 months?"

Table 1. The above table lists all risk-determining sections and a sample question in each section in the YSET questionnaire.

Dataset. From the questionnaire dataset we received directly from the GRYD program, we observed a total of 32,896 responses from 22,567 unique participants, with multiple measurements of intakes and retakes from 2008 to 2019. In addition to the questionnaire results, each response contains temporal and locational (registered GRYD district) information of the response as well as demographic information regarding the participant (such as age, gender, and race). Despite the large number of participants and responses, we only identified 4,030 participants with complete responses who qualified for the secondary program and have been enrolled in the program for at least half a year, 1,663 of which have been enrolled in the program for at least a full year. This scarcity of observations is due to the fact that a large proportion (65.67%) of the participants did not qualify for the secondary program, graduated from the program, or dropped out of the program in the first half-year period.

It is important to note that questions are scaled differently across the questionnaire. Some questions are scored on a scale of 1 to 5 or on a scale of 1 to 4, while others are binary. For the purposes of our analyses, we scaled the questions we take into consideration so that they all fall between 0 and 1.

Areas of Interest. For this project, we have focused on three main areas of interest. The first is to evaluate the change in responses of individual questions as participants move through the GRYD program over time. Such results would be able to provide insight as to which topics these questions cover the GRYD program may be effectively or ineffectively addressing. Next, we would like to evaluate the importance of question responses at one time in calculating a future risk score, meaning the risk score a participant receives after they have been receiving services over a period of 6 months. The goal is to give GRYD a better understanding of the "high-risk" labels they are assigning to certain participants. Lastly, we investigate making predictions on future questionnaire responses and on other risk indicating scores. To investigate this, we narrowed our data to focus on the subset of 1,663 participants who had at least 3 responses (corresponding to 2 GRYD program periods, Y1-R1 and R1-R2) recorded throughout at least a full year of being enrolled in the GRYD program. This was the largest data subset with multiple time snapshots, which we needed to analyze the dynamics of participants' questionnaire responses over time. In addition, we focus our analyses on the 39 attitudinal questions, as there was concern for inaccurate responses recorded in the behavioral questions. As we looked into our areas of interest, we also came across the possibility that differences in demographics could affect our results, and so we do include analyses on these different groups.

The paper is outlined as follows. In 'Model Building and Analysis', two dynamical models are built and analyzed in order to extract the dynamic relationship behind the YSET data. This is followed by the verification of these two methods by examining their prediction accuracy and by comparing them with machine learning algorithms. Lastly, we conclude with an overview of the results and a discussion of future work.

Model Building and Analysis

Dynamic Mode Decomposition (DMD). Dynamic Mode Decomposition (DMD) is a method that analyzes the approximate linear dynamics of a system over time (2). Though it has traditionally been employed within the field of fluid dynamics (3), we were able to apply it on our data by interpreting each participant's response and risk score as a vector in a corresponding "participant space" and using the approximation of the governing equation to understand how the GRYD program affects each individual. The DMD algorithm requires a known input and output matrix, X and Y. In the case of our data, these matrices consist of the 39 attitudinal question responses. The input matrix consists of the participants' question responses at one time and the output matrix consists of those at the next time. Using these matrices, the algorithm produces a linear transformation matrix A, whose eigendecomposition reveals the system's growth or decay and spectrum.

Algorithm. The dynamical system is represented by

$$AX = Y \tag{1}$$

, where X and Y are known $m \times n$ matrices and A is an unknown $m \times m$ matrix. A is solved by the Least Squares solution:

$$A = YX^{+} \tag{2}$$

. To begin, we compute the Singular Value Decomposition (SVD) of A:

$$A = U\Sigma V^{*} \tag{3}$$

, where $U \in \mathbb{C}^{m*m}, \tilde{\Sigma} \in \mathbb{C}^{m*m}, \tilde{V*} \in \mathbb{C}^{m*n}$. With this, we directly compute the transformation matrix A:

$$A = YX^{+} = YV\Sigma^{-1}U^{*} \tag{4}$$

Interpreting change in responses.. The eigendecomposition of the transformation matrix A provides a list of eigenvalues and their associated eigenvectors, among which we can obtain and analyze the dominant eigenvalue and its associated eigenvector (dominant eigenvector), as they contribute the largest weight in any linear combination representation of a participant's response. The dominant eigenvalue indicates the overall trend of the system: if the magnitude is greater than 1, it means the system is growing; while less than 1, the system is decaying. To measure each question's susceptibility to change, we define *Change* to be the magnitude of the dominant eigenvector's entries, since they allow us to understand the growth or decay of responses to individual questions and risk scores over time. We calculated transformation matrices A for both periods (intake to retest1 and retest1 to retest2) and did the corresponding eigendecomposition. The results of eigenvalues and the dominant eigenvector are plotted as shown in Figures 1 - 3.

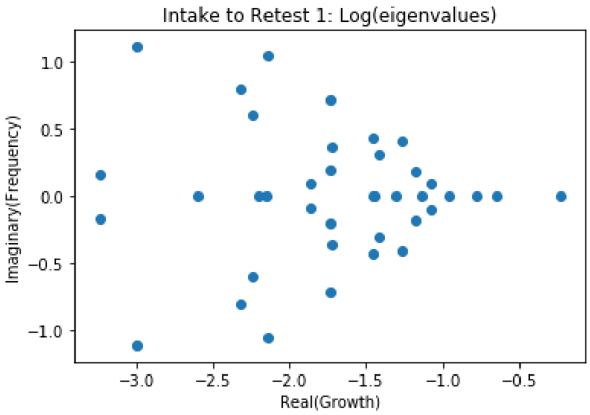

Fig. 1. Log of eigenvalues from Y1 to R1. All logs are smaller than 0, which means which means that the magnitude of all eigenvalues is smaller than 1. The system of question responses is decaying in period Y1 to R1. This suggests that GRYD program is effective in lowering risk scores on the YSET in period Y1 to R1, which could be an indicator that the participants would be less likely to join a gang.

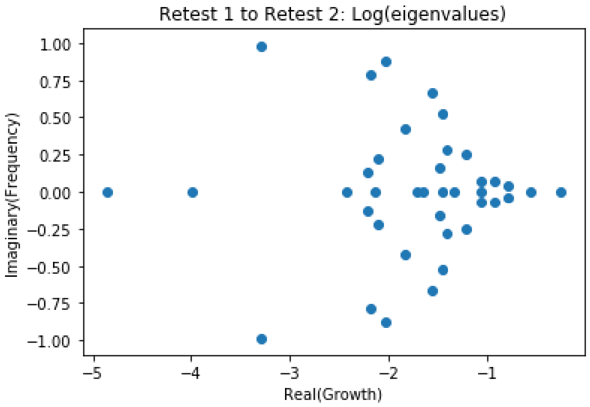

Fig. 2. Log of eigenvalues from R1 to R2. All logs are smaller than 0, which means which means that the magnitude of all eigenvalues is smaller than 1. The system of question responses is decaying in period R1 to R2. This suggests that GRYD program is effective in lowering risk scores on the YSET in period R1 to R2, which could be an indicator that the participants would be less likely to join a gang.

Figures 1 and 2 show that log of eigenvalues in both periods are smaller than 0, which means that the magnitude of all eigenvalues is smaller than 1 in both time periods. From our interpretation, the system of question responses is decaying, and the scores to survey questions are lower. In other words, this suggests that the GRYD program is effective in lowering risk scores on the YSET, which could be an indicator that the participants would be less likely to join a gang.

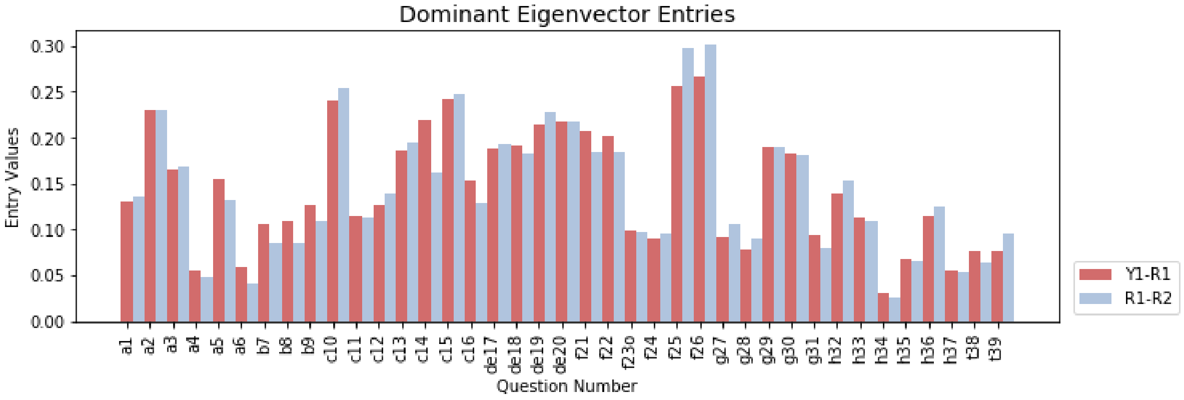

Fig. 3. Juxtaposition of dominant eigenvector entries from Y1 to R1 and R1 to R2. Red represents each question's susceptibility to change in period from Y1 to R1, and blue represents that from R1 to R2. Questions with the highest entry values are most reluctant to change, while those with the lowest entry values are most susceptible to change. Questions f25 and f26 are most reluctant to change, while question h34 is most prone to change.

Figure 3 shows each question's susceptibility to change indicated by its corresponding entry value in the dominant eigenvector. Questions with the highest entry values are most reluctant to change. The questions with the highest two entry values are f25 (It is okay to beat people if they beat me first) and f26 (It is okay to beat people if they beat me up). Questions with the lowest entry values are most susceptible to change. The question with the lowest entry value is h34 (How many of your friends have attacked someone with a weapon?)

Importance of questions in determining future risk score. Apart from analyzing the eigendecomposition of the linear transformation matrix A produced by the DMD algorithm to observe the change in question responses over time, we developed a method in which we can use DMD to observe the importance of participants' question responses at one time in calculating their future risk score. We assume the participant's future risk score can be calculated as a linear combination of the participants' question responses from a previous time, and in fact, DMD is able to provide us with the coefficients in this linear combination. It is the weights of these coefficients that then reveal whether certain question responses are more important or not. To produce these results using DMD, we manipulated our input and output matrices to include the participants risk scores assigned at the associated time in addition to their attitudinal question responses. This was appended as a another row to these input and output matrices. The DMD algorithm was then executed as before with these altered matrices, producing another linear transformation matrix A. Figure 5 displays a schematic of the method.

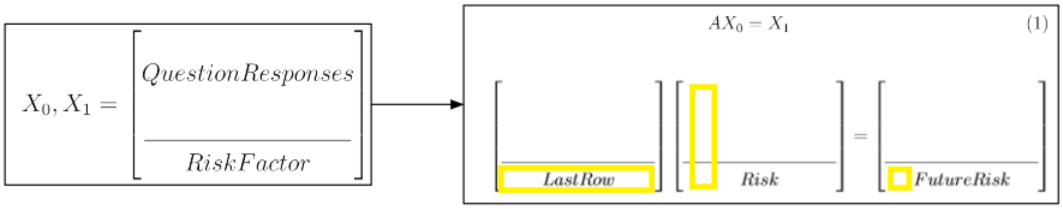

Fig. 4. Schematic of the modified DMD inputs to calculate question importance

Instead of analyzing its eigendecomposition, we look at the last row of the matrix A to observe the weights of the coefficients of the question responses in the linear combination of calculating the future risk score. The weights of those found in both linear transformation matrices A, from Y1 to R1 and R1 to R2, are presented in figure 5.

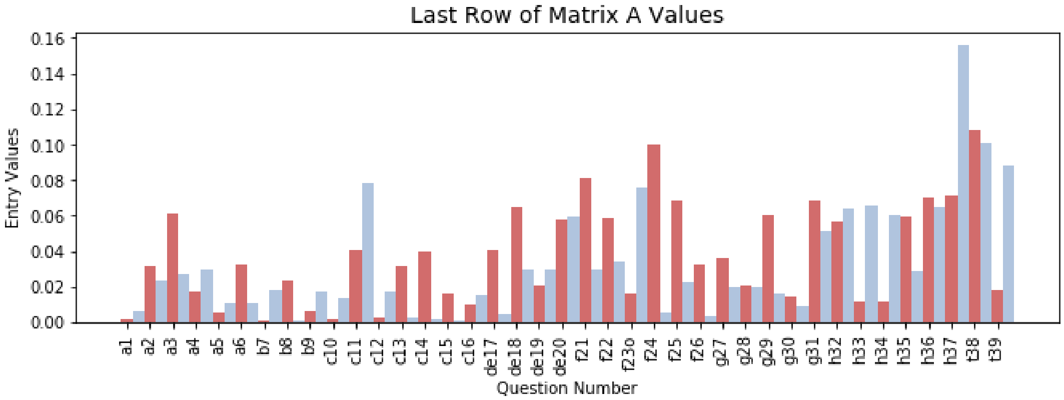

Fig. 5. Juxtaposition of weights of entries in the last row of matrix A from Y1 to R1 and R1 to R2. Red represents the importance of each question in determining the future risk from Y1 to R1, and blue represents that from R1 to R2. Questions with higher entry values are more important in calculating the future risk score, while those with the lower entry values are less important.

Change-Importance Coordinates. As a reminder, we define *Change* to be the magnitude of the dominant eigenvector's entries, and *Importance* to be the magnitude of entries in the last row of the transformation matrix. After examining change and importance in previous sections, we wanted to understand the relationship between change and importance. Here, we will introduce the change versus importance in figure 6.

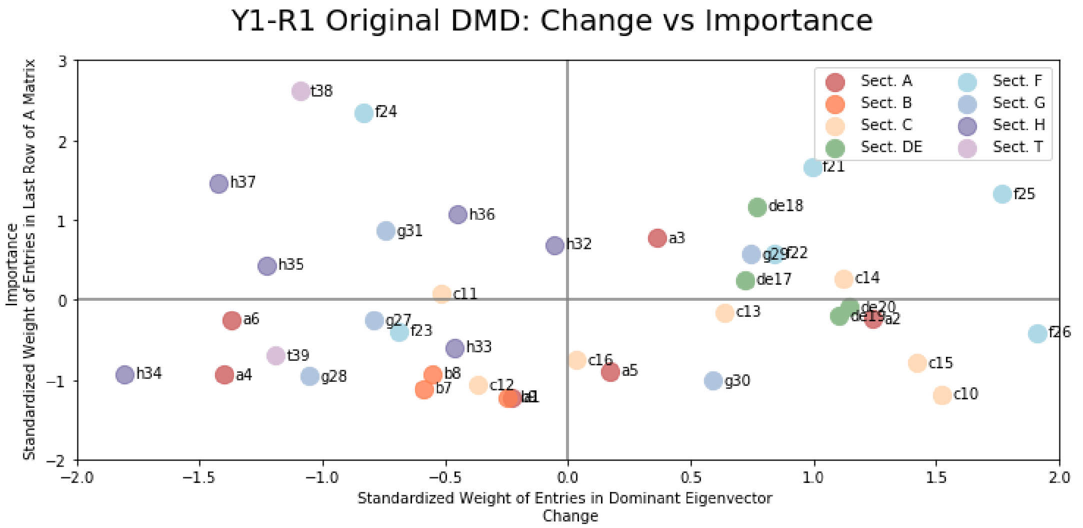

Fig. 6. Plot between change and importance. Horizontal axis represents change, and vertical axis represents importance. Each color represents one section. The graph shows the relationship between change and importance for each question. Questions in the upper right quadrants are important in determining future risk scores yet reluctant to change, while questions in the lower left quadrants are less important and more susceptible to change.

In Figure 6, the horizontal axis represents change, and vertical axis represents importance. Z-scores of the corresponding entry in the dominant eigenvector and that in the last row transformation matrix A are calculated for each question as a coordinate pair. The graph shows the relationship between change and importance for each question.

Shortcomings of DMD. Though DMD can be used to analyze the dynamic relationship behind YSET data, it has some shortcomings. DMD can not explain the variation in change among different characteristics of participants (such as gender). Figure 7 shows the importance of each question for female and male participants in both periods. The distinct patterns between genders suggests that it may not be appropriate to treat females and males using one general DMD. Therefore, a variant of DMD method is introduced below.

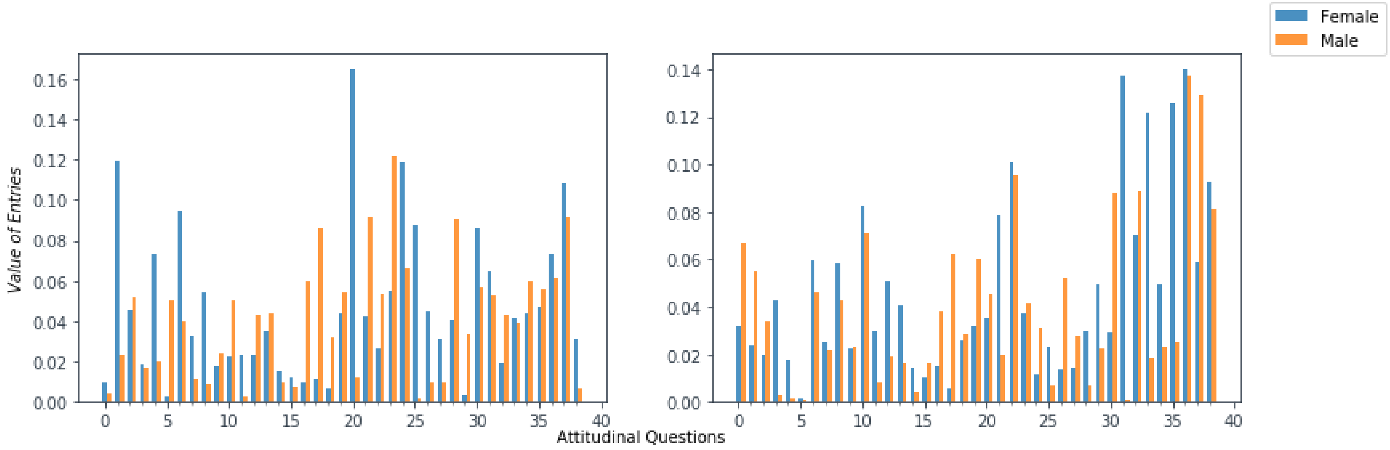

Fig. 7. Plot of entries in the last row of the transformation matrix A from Y1 to R1 and R1 to R2. Blue represents males, and orange represents females. Left graph represents importance of each question in determining future risk from Y1 to R1, and right represents that from R1 to R2. Almost all questions have different importance weights for different genders, so treating gender using a general DMD may not be appropriate.

Dynamic Mode Decomposition with Control(DMDc). (4) Dynamic Mode Decomposition with control (DMDc) is a modified version of DMD (5). The goal of DMDc is to analyze the relationship between a future system measurement X_{k+1} with the current measurement X_k and a current control C_k. The mathematical formula is

$$X_{k+1} = AX_k + BC_k. \qquad [5]$$

Similar to the original DMD, DMDc is typically utilized to investigate physical dynamical systems, where C_k representing external control obeys physical laws. In the context of GRYD, the matrix X_k and X_{k+1} are same as those used in DMD. We used demographic information of each participant as the control factor to the system, so that DMDc can help us discover the underlying dynamics without the confounding effect of demographic differences between participants.

Algorithm. The dynamical system is represented by

$$AX_k + BC_k = X_{k+1} \tag{6}$$

X_k and X_{k+1} are two known $m \times n$ matrices, and C_k is a known $p \times n$ matrix; A is an unknown $m \times m$ matrix, and B is an unknown $m \times p$ matrix. A and B are solved by the Least Squares solution. To begin, we construct a $(m + p) \times n$ matrix Ω by combining X_k and C. Then we compute the Singular Value Decomposition (SVD) of Ω:

$$\Omega = U \Sigma V^*, \tag{7}$$

where $U \in \mathbb{C}^{(m+p)*(m+p)}, \tilde{\Sigma} \in \mathbb{C}^{(m+p)*(m+p)}, \tilde{V^*} \in \mathbb{C}^{(m+p)*n}$. After that, we split U into two parts, top as U_1 and bottom as U_2. With these two matrices, we directly compute matrix A and B using U_1 and U_2 respectively:

$$A = X_{k+1} V \Sigma^{-1} {U_1}^* \tag{8}$$

and

$$B = C_{k+1} V \Sigma^{-1} {U_2}^* \tag{9}$$

We used the following control factors.

Age Control. The age of participants is distributed to 7 different categories, from 10 years old to 16 years old. Therefore, our control (indicator) vector has length 7. In total, 1662 samples are available, training set and testing set contains 1330 and 332 samples respectively.

District Control. The GRYD participants in this dataset are located in 21 GRYD zones in total by integers from 1 to 23 except for 6 and 16. Therefore, the control vector has length 21. Also, as we can not make sure that whether people will move throughout the period, we have to drop all data that has a missing district. Therefore, 1632 samples are left, in which 1306 are in training set, 326 are in testing set.

Ethnicity Control. The GRYD program divides the ethnicity to 5 categories: Asian, Black, Latino, White, and Other, so the control vector has length 5. Also, some children report that they have multi-ethnicity, and we include all their possible ethnicities. However, for those whose responses are missing, we drop the data. Our data set contains 1653 samples, in which 1322 samples are in training set, and 331 samples are in testing set.

Gender Control. We use $(1, 0)$ to denote male and $(0, 1)$ to denote female. If the report of the child is not consistent (that is, reported male in Y1 and female in R1), we choose to use the gender reported in the Y1 data set. The data set contains 1663 samples, in which training set contains 1330 samples and testing set contains 333 samples.

Analysis. From the results obtained from all the above mentioned models, we were able to analyze the change in response and importance in determining future risk scores of each question from different perspectives. By observing each question's change-importance coordinates, we interpreted the different patterns resulted from different models from both an within-period and between-period point of view while keeping in mind of the potential demographic effects.

Difference in Question Behaviors Within Program Periods. Within the same program periods and in all models, we observed that most question responses from Section F, H and T (addressing risk-influencing factors of Neutralization, Peer Delinquency and Family Gang Influence, respectively) tend to be more important in determining the risk score of an individual measured at the end of the current period (i.e., that of the beginning of the next period). Among these sections, Section F (Neutralization) is more prone to change than the other two sections. Since the total risk scores of all participants are generally decreasing, we infer that the change in response as a numerical decay, meaning that the GRYD program is effective in lowering the risks in the corresponding risk-influencing areas.

On the other hand, Section DE (Impulsive Risk Taking) is more reluctant to change in all models within the same program period. Such reluctance suggests that the GRYD program did not lower risk scores in this area as much as in other areas. Since questions in this section also fall in the more important quadrants as seen in Figure 8, to further improve the effectiveness of program, we suggest GRYD to consider targeting services to help reduce in participants' impulsiveness during the program periods.

Difference in Question Behaviors Between Program Periods. As seen in Figure 9, although a lot of the questions behave similarly in most models and both program periods, there are still questions that behavior very differently for different models and different time periods. It is worth noting that, despite the clear difference between the results produced by DMD and DMDc models, many questions demonstrate a significant decay in relative importance from the first program period to the second (as seen on the right side of Figure 9).

We also found that while responses to questions in Section H (Peer Delinquency) seem to be decaying in the first program period, they become reluctant to change in the second. This trend suggests that if the program works the same way in both time periods, the effectiveness in lowering risk in the peer delinquency area might decrease over time.

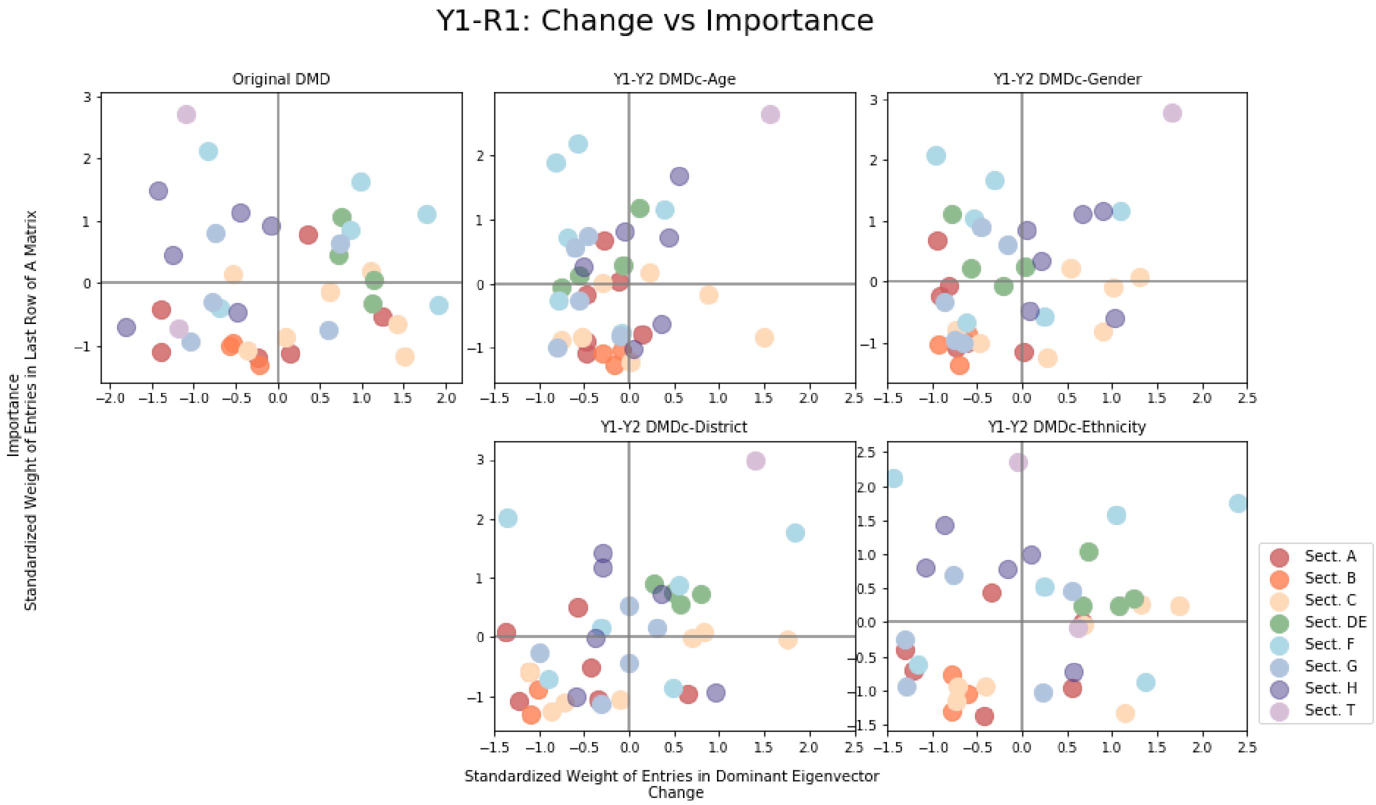

Fig. 8. The above figures plot the change-importance coordinates for each question during the first program period (between the questionnaire responses Y1, the intake, and R1, the first retake) for all DMD and DMDc models. The dots are color coded by sections. We can see that Section H in purple and Section F in sky blue are more important relative to other sections, while Section DE in green and Section C in yellow are more reluctant to change in all 5 models.

Difference in Question Behaviors Between Demographic Groups. Another trend we find in Figure 9 is that in questions like A3, F21, H33 and H35, the change-importance coordinates for the original DMD is clearly distinct from all of the DMDc models, and the difference is especially significant in their reluctance to change. This trend is suggesting that different demographic groups have different behaviors in answering these questions, so it might be more accurate to construct different dynamical systems for each demographic group.

Difference in Question Behaviors Within Sections. Questions in this questionnaire are grouped into sections that reflect the same risk-influencing factors, so one might expect that questions within each section might have similar behaviors across models and time. While this holds true for most of the sections (especially Section B and DE, corresponding to areas of Weak Parental Supervision and Impulsive Risk Taking), questions in Section F (Neutralization) and T (Family Gang Influence) have very different behaviors even in the same model under the same program period.

One explanation for this difference in questions could be the phrasing of the questions. For example, Section T contains the following questions that both address the risk of family gang influence:

- T38: How many people in your family think that you will join a gang?
- T39: How many people in your family are gang members?

From Figure 10, we can see that T38 is important in determining the future risk score in both program periods, yet T39 is only important in the second program period. By examining the content of the questions, we found that T38 is asking about the potential of family members joining a gang (i.e., more attitudinal or opinion-based), while T39 is asking for an objective numerical response. It is possible that T39 does not play such a significant role in determining the future risk score if participants are more cautious when responding to their first questionnaire and thus might not answer these questions truthfully.

Verification on Model Validity

DMD and DMDc comparison. It can be shown that the DMD algorithm we applied minimizes the square error of prediction (see Lemma 2 and Lemma 3 in the supplementary materials). Similarly, the DMDc algorithm we applied minimizes the square error of prediction (Lemma 3 and Lemma 4 in the supplementary materials). In the next section, we confirm this for our numerics.

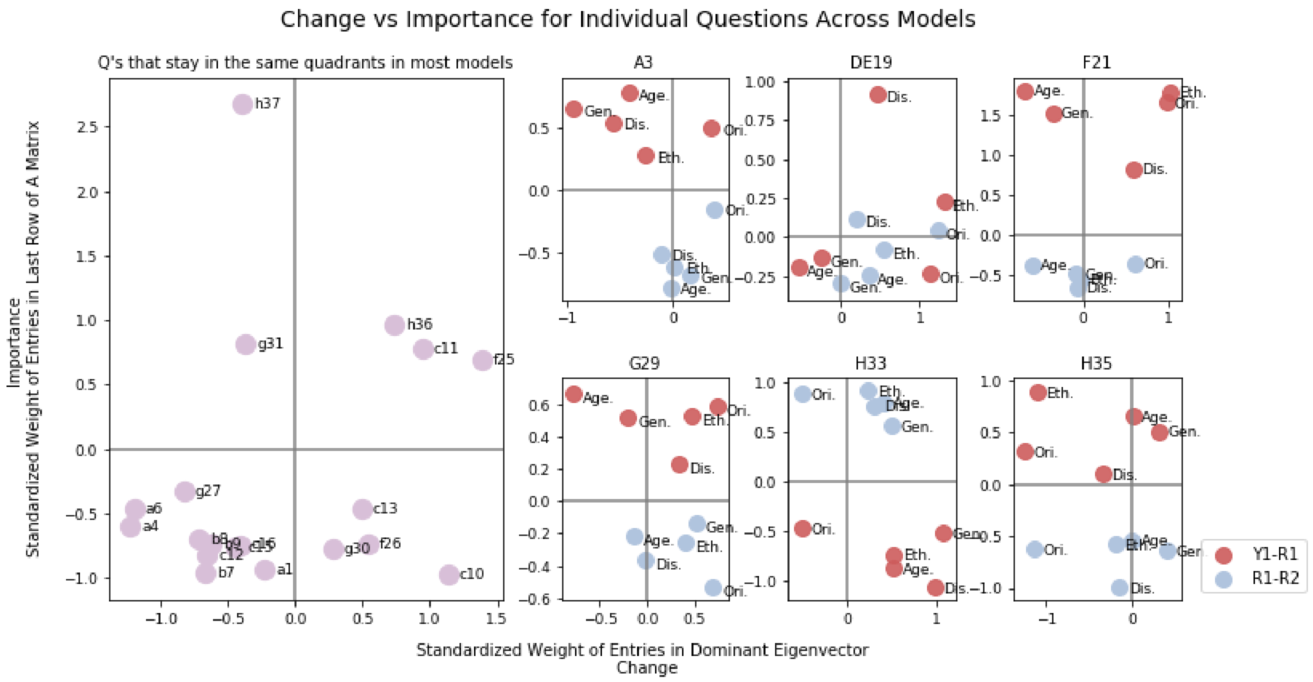

Fig. 9. The above figure shows different question behaviors in different models. The plot on the left contains questions that stay in 70% or more of all the models in the two program periods, and the coordinate of each question is determined by averaging its standardized change and importance scores respectively across all models. The plots on the right are individual questions that behave differently for different models. Each dot in a subplot represents the corresponding question's behavior in one of the models, and the color represents the program period. We can see a significant difference in importance between the two program periods, and for many questions there is a difference in their reluctance-to-change value between the original DMD model and the DMDc models. These trends show that we need to take into consideration of the effect of demographic information as well as the time period when evaluating each question's behavior.

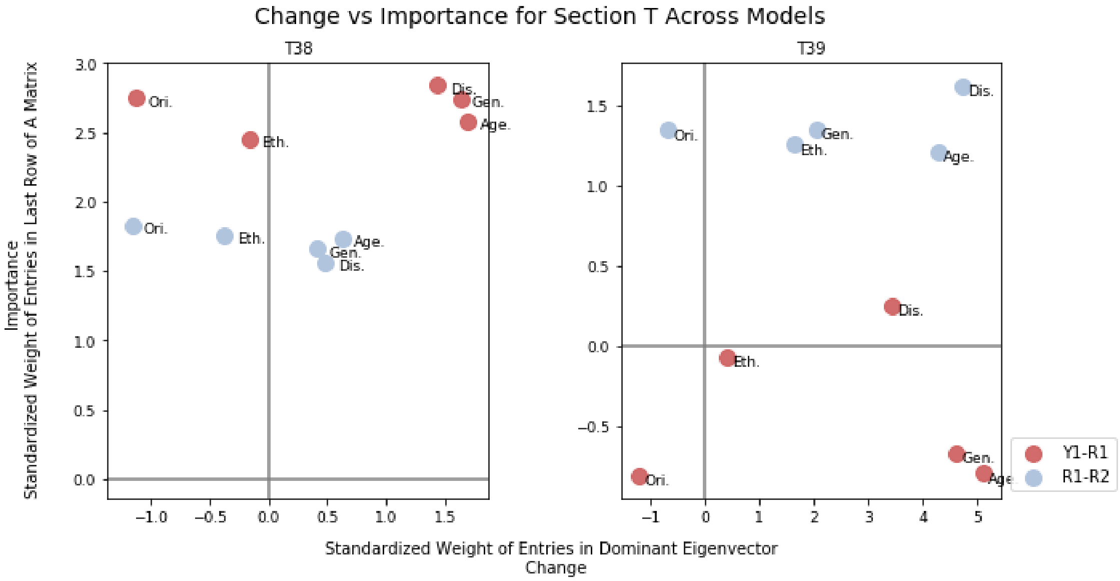

Fig. 10. The above figure plots the change-importance coordinates for both questions in Section T in all models. The red color represents the first program period, and the blue color represents the second. We can see that T38, a more opinion-based question, has a high relative importance when determining future risk for both program periods, while the more fact-based T39 is only significant during the second program period. This difference suggests that different ways of question phrasing in the same risk-influencing factor can result in different behavior of question responses as well.

Prediction & Errors. We randomly split the data into 80% training set and 20% testing set in each trial. We then calculated the mean square error between the prediction and the actual data for validating the accuracy. We implemented the DMD without control and with four different controls (age, district, ethnicity, and gender respectively), running the trial 100 times each, to calculate the mean square error between prediction and real data. Additionally for comparison purposes, we also used DMD and DMDc model trained by 80% data to predict the entire data set and calculated the mean square error. The results of 5 sample trails over DMD without control and with four controls are shown in Table 2.

DMD (No Control)		
Trial	MSE (Test)	MSE (Whole)
1	0.1018	0.0965
2	0.1004	0.0964
3	0.1012	0.0964
4	0.1006	0.0964
5	0.0993	0.0964

DMDc (Age)		
Trial	MSE (Test)	MSE (Whole)
1	0.0981	0.0948
2	0.0983	0.0949
3	0.1025	0.0949
4	0.1009	0.0950
5	0.1020	0.0950

DMDc (District)		
Trial	MSE (Test)	MSE (Whole)
1	0.0988	0.0924
2	0.1009	0.0924
3	0.1007	0.0924
4	0.1028	0.0923
5	0.1019	0.0924

DMDc (Ethnicity)		
Trial	MSE (Test)	MSE (Whole)
1	0.0990	0.0956
2	0.1030	0.0956
3	0.1028	0.0957
4	0.0995	0.0957
5	0.1025	0.0957

DMDc (Gender)		
Trial	MSE (Test)	MSE (Whole)
1	0.1027	0.0957
2	0.1014	0.0956
3	0.1007	0.0956
4	0.0999	0.0956
5	0.0988	0.0956

Table 2. Sample MSE of DMD and DMD with control trials: these trials are sample trials picked from 100 trials with 80% of the data as training data and 20% of the data as testing data. The MSE is calculated based on the models performance on the testing set and the whole set. The previous one measures the robustness of the model, while the latter one measures the overall performance of the model on the dataset. From the table, we can observe that in general, DMD with control has a lower MSE than DMD.

Significance in mean-squared error between DMD and DMDc.. In order to see whether the difference between DMDc and DMD is significant (that is, if DMDc significantly decreases the MSE compared to DMD), we need to implement a t-test. Our samples are composed of the MSE from DMD and DMDc on different controls for 100 trials each. Suppose the means are denoted by μ and μ_c. Our hypothesis is shown in Table 3 (where H_0 represents the null hypothesis).

$$H_0 : \ \mu = \mu_c$$

$$H_a : \ \mu \neq \mu_c$$

DMD and DMDc (Age)		
MSE	t statistic	p value
Test	4.1337	0.0001
Whole	232.4665	0.0000

DMD and DMDc (District)		
MSE	t statistic	p value
Test	6.9643	0.0000
Whole	403.1224	0.0000

DMD and DMDc (Ethnicity)		
MSE	t statistic	p value
test	0.4925	0.6229
Whole	121.0410	0.0000

DMD and DMDc (Gender)		
MSE	t statistic	p value
Test	4.4854	0.0000
Whole	148.0830	0.0000

Table 3. t-test of DMD and DMD with control: in order to verify our assumption that DMD with control has better performance, using those 100 trials, we calculate the mean and variance of the MSE and perform a t-test to see whether the results from two algorithms have the same mean. The null hypothesis is that the MSE from DMD and DMD with control have the same mean. From the result, we can observe that, with the exception of the ethnicity control group, all other t-tests reject the null hypothesis. This means that the MSE mean of DMD with control is different from the MSE mean of DMD. Also, from the t-score, we can conclude that the MSE mean of DMD with control is smaller than that of DMD.

Analysis of MSE and t-test. In line with the theoretical results, the DMD and DMDc always generates the least square solution, and in DMDc, we added an additional matrix B for prediction, the mean square error of any DMDc should always be smaller than DMD when the entire data set is used for both training and prediction. However, since we are splitting the data into training set and testing set randomly each time with ratio of $4:1$, there are some variations when applying DMD and DMDc in each trial, but the DMDc generally performs better than DMD in the sense of minimizing the mean square error.

Among the four controls we added, the control on district generates better prediction than other controls, especially on the prediction of the whole set. This suggests that the pattern of variation due to different district is clear, and different approaches in the youth development programs should be applied based on the different districts. However, DMD with control on ethnicity does not improve the prediction much according to the low t-statistic and shows greater variation among different trials. There are two reason for the low t-statistic and high variation. Firstly, there might be no apparent pattern of difference among different ethnicity of people, and secondly, a lot of people declared multi-ethnicity, and all these people are categorized into "others", thus influencing the accuracy of the prediction. The control on age and gender also perform significantly better result than the original DMD, suggesting that the age and gender can explain a relatively large proportion of youth's change between two periods. However, some of the categories only include a small number of samples, which means that when splitting the training and testing data, if the samples of these categories are not included in the training set, it may result in the bias and imprecision in prediction.

Machine Learning Methods to Validate DMD Results

Validation of Question Importance. Though DMD creates seemingly reasonable results, we need evidence to confirm the validity of the DMD method. That is, if we implement other machine learning methods, we want to verify that we obtain similar results to the DMD results. In this section, we will verify the importance of questions. In our analysis, the contribution of each question to the future risk factor is determined by the coefficients of the last row of the transformation matrix A. We can use machine learning techniques like linear SVM or decision tree to show the importance of different features.

Support Vector Machine (SVM) is a well-known classification technique. The method aims to use a hyperplane (i.e., the support vector) to separate data samples into different groups. If the data is not separable by the support vector in lower dimensions, SVM will use a kernel function to project the data to a higher, linear-separable dimension. The absolute value of the coefficient of the support vector represents the importance of each corresponding dimension. More introduction about SVM is included in the supplemental material (6). Decision Tree (DT) is another classification technique that clusters data samples in a tree-like structure. The tree starts with a root node containing all the data and classifies data into branches by different features in each node. The tree grows as more branches are created, and it will eventually produce a tree whose leaf nodes represent samples in the same cluster. The choice of the feature is based on loss functions called "Gini loss" or "entropy loss." With this metric, we can evaluate the importance of features based on their contributions to the information gain of the tree. More introduction about Decision Tree is included in the supplemental material (7).

In each trial, we scale the data into the range of $[0, 1]$ for SVM. Since SVM is mainly used for binary label classification, we change the label of data from risk factor to eligibility (1 when risk factor greater than 3, 0 when risk factor smaller or equal to 3). For decision tree, though there is no need to re-scale the data, we can not directly use the feature importance from the decision tree as the real importance of each questions. As the "feature importance" in the decision tree is to compare the total information gain from each question, different questions are used sequentially instead of in parallel as criterion. Therefore, we need to revise the process of decision tree as we describe in the following paragraph.

Our decision tree only contains 1 decision node, which is the root node. We iterate over the features (questions). At each time we iterate over the features (questions), we use only one question and compare to see which threshold produces the most information gain. The importance of questions can be represented by the maximum information gain using that question as a criterion.

We focus on verifying the DMD model from Y1 questionnaire to R1 questionnaire. The following figures show the question importance derived by the SVM method and decision tree method. To show the data clearly we re-scale the information gain of decision tree from 0 to 1.

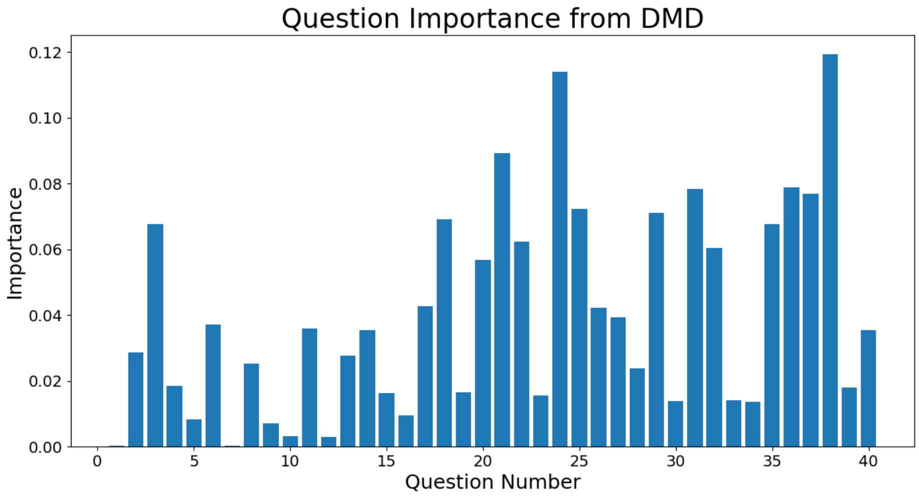

Fig. 11. Question Importance from DMD Algorithm: the columns indexed from 1 to 39 stand for 39 questions, and the last column stands for the importance of risk factor. The figure is derived from the last row of transformation matrix A, in order to show the importance of each question when predicting future risk factor. In this figure, we can observe that questions indexed 3,21,24,38 have relatively high importance.

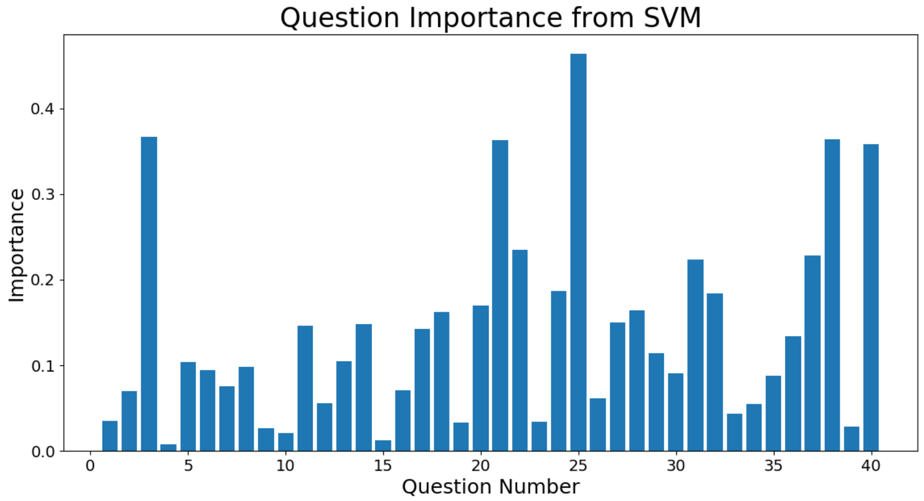

Fig. 12. Question Importance from SVM: the columns indexed from 1 to 39 stand for 39 questions, and the last column stands for the importance of risk factor. The figure is derived from the absolute value of the support vector from linear SVM. The absolute value of the support vector stands for the importance of each question in classification. In this figure, we can observe that question 3,21,24,38 have relatively high importance as well, consistent with the result from DMD.

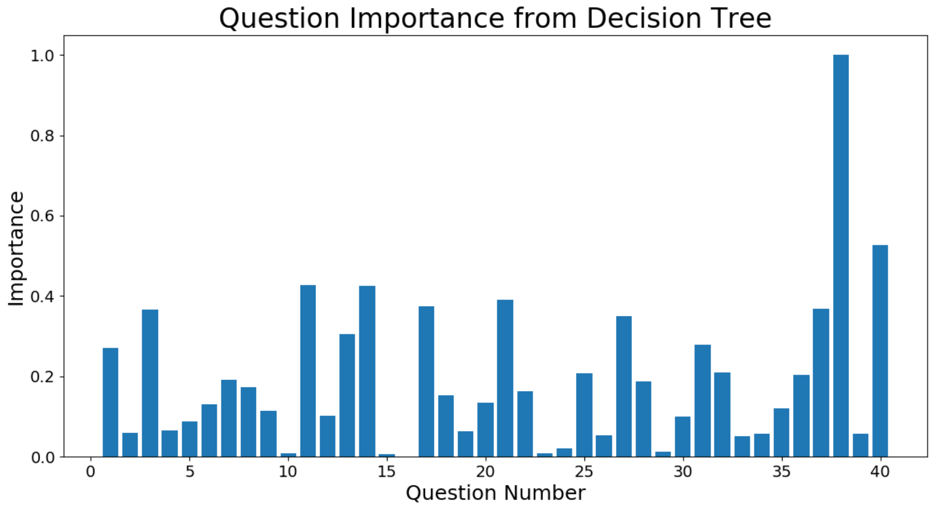

Fig. 13. Question Importance from Decision Tree: the columns indexed from 1 to 39 stand for 39 questions, and the last column stands for the importance of risk factor. The figure is derived from the information gain using each question as criterion to classify the data. Higher information gain stands for higher importance of the question. In order to maintain the same magnitude, we scale the result to the range 0 to 1. In this figure, we can observe that questions 3,11,14,21,38, and the risk factor have relatively high importance.

We observe similar importance scores using DMD and SVM. Although the result from DT is quantitatively different, it provides the same question index for the highest importance score (see figure 11, 12, and 13). We also calculated the correlation of the importance scores from DMD, SVM, and DT methods. The correlation coefficient matrix is shown in Table 4.

	DMD	**Linear SVM**	**Decision Tree**
DMD	1	0.6823	0.4675
Linear SVM	0.6823	1	0.5463
Decision Tree	0.4675	0.5463	1

Table 4. Correlation coefficient matrix of the question importance of three algorithms. The matrix represents the correlation of the importance of each question evaluating using three different algorithms. Higher correlation means that two algorithm give more similar results. We use this matrix to verify that our analysis of question importance using the DMD algorithm is valid.

The correlation coefficient matrix in table 4 shows that importance scores of the three methods are correlated. In particular, we observe strong correlation between the results from DMD and linear SVM. We also perform the Spearman's Ranking-Order Correlation test to see whether the questions that have higher importance in one model will also have higher importance in the other models. The null hypothesis of the test is that two sets of data are uncorrelated, while the alternative hypothesis is that the two sets are correlated.

The Spearman's Ranking-Order Correlation matrix is shown in table 5, and the test result is shown in Table 6.

	DMD	Linear SVM	Decision Tree
DMD	1	0.6456	0.4315
Linear SVM	0.6456	1	0.5510
Decision Tree	0.4315	0.5510	1

Table 5. Spearman's ranking-order correlation matrix of three methods. Spearman's ranking-order measures that if the data is sorted, according to the rank, whether samples in two datasets with similar numerical ranking have similar index. That is, higher correlation means the ranking distributions of two datasets are more similar.

Pair of Models	p-value
DMD & Linear SVM	6.85×10^{-6}
DMD & Decision Tree	0.00543
Linear SVM & Decision Tree	0.00023

Table 6. Spearman's ranking-order correlation test result of the three algorithms. This test is used to measure the significance of Spearman's ranking-order correlation. The null hypothesis is that two sets of data are uncorrelated. From our trials, we can observe that the null hypothesis is rejected for all three comparisons under the significance level of 1%, which show that the three sets of question importance are actually correlated.

We observe that all three null hypotheses that two sets of data are not correlated are rejected under 1% significance level. Therefore, we can conclude that the DMD, linear SVM, and decision tree results are consistent, which supports our use of the DMD method to analyze the importance of questions.

Future Work

Comparison of DMD and Neural Network. A neural network is a structure that can theoretically approximate any governing function of systems. The network contains three parts: an input layer, hidden layers, and an output layer. All nodes are fully connected by linear functions of the form $w \cdot x + b$. Hidden nodes and output nodes can be activated using a nonlinear activation function. The training process of the network is divided into two parts: a feed forward process and a back propagation process. Further explanation of neural network is included in the supplemental material (6).

Based on the structure of the neural network, we notice that if we remove the activation layer of the neural network, then the structure of the neural network will become linear. Therefore, it is possible for the linear neural network to simulate the result of DMD using different approaches (stochastic gradient descent and other related algorithms).

Further Validation of the DMD Model. To confirm that our DMD model can provide a suitable prediction of the R1 questionnaire from the Y1 questionnaire, we can implement an alternative method to approximate the optimal matrix A^* and check whether those two matrices are similar or have similar properties. One possible approach is to use Stochastic Gradient Descent (SGD). SGD is an iterative method that approaches the minimal of an objective function by moving along the direction of negative gradient. In order to implement SGD in a fast speed and preserve the linear structure of the model, we can use neural network without the activation layer. The neural network without the activation layer is a linear transformation (see supplementary material). We implement the neural network without activation layer using the platform Keras based on Tensorflow.

To check the result of DMD, we construct a neural network with 40 inputs (39 questions and 1 risk factor) standing for questionnaire of Y1. The corresponding output layer has 40 nodes, and the hidden layer has 256 nodes. The loss function we use is the Mean Squared Error (MSE), and the optimization algorithm we use is called "Adam", which is an algorithm with an adaptive learning rate for the gradient descent. Using the algorithm "Adam", we can try to avoid the network from falling into local minimum, which is one of the drawbacks of gradient descent algorithms. After training of the neural network, as the structure is linear, we can derive the transformation matrix A_{NN} and bias b from the network. The result of the neural network is supposed to be close to the global minimum (small error may occur based on number of epochs, learning rate, and other hyper parameters). Some theoretical proofs of the error bound of DMD and neural network is included in the supplemental materials. We repeat the DMD trial and neural network trial using 80%:20% train-test split for 100 trials. The mean squared error (MSE) of the test set and whole dataset of the DMD and neural network is shown in table 7.

The two methods provide us with similar MSE. Consistent with the proof in the support material, the error of neural network is slightly bigger than that of DMD, which means that DMD is stable and provides us with good results. According to our simulations, the lower bound of MSE (using the whole dataset as training set) of DMD is approximately 0.9582. Furthermore, we need to compare whether the two matrices A_{DMD}, A_{NN} are similar by comparing their eigenvalues and singular values. The comparison of two values are shown in figure 14.

DMD			NN		
Trial	MSE (Test)	MSE (Whole)	Trial	MSE (Test)	MSE (Whole)
1	0.1007	0.0965	1	0.1000	0.0966
2	0.1010	0.0964	2	0.0989	0.0967
3	0.1018	0.0964	3	0.1007	0.0968
4	0.1033	0.0964	4	0.1002	0.0973
5	0.1004	0.0964	5	0.1011	0.0967
Mean	0.1013	0.0964	Mean	0.1018	0.0975
std	0.00169	0.00041	std	0.00195	0.00049

Table 7. MSE comparison of DMD algorithm and linear neural network. The first five trials are sample results picked from 100 trials. As we randomly pick the training set and testing set, each of the 100 trials provide a different result. Therefore, we calculate the mean MSE and standard deviation for those 100 trials. From the result, we can see that those two methods provide similar MSE values.

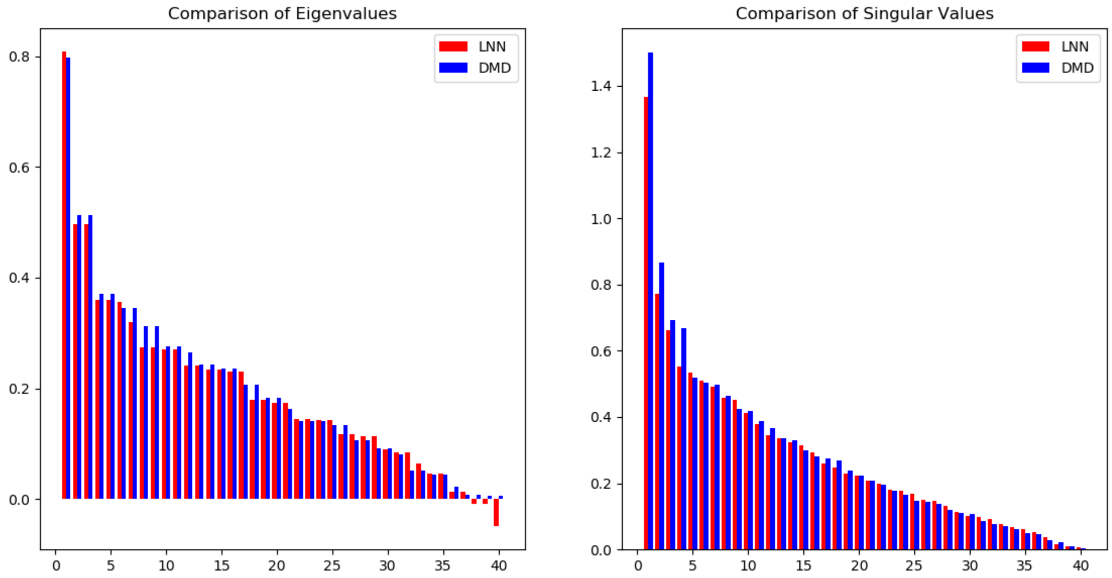

Fig. 14. Comparison of eigenvalues and singular values of DMD and linear neural network: the eigenvalues and singular values are sorted in decreasing order, and only real part of both values are picked. The red bars stands for the result of linear neural network, and the blue bars stands for the result of DMD. From the figure, we can observe that the results of two methods have similar distribution and values with correlation coefficient close to 1. It shows that the two matrices have similar properties.

From Figure 14, we can observe that for all eigenvalues and singular values, A_{DMD} and A_{NN} provide us with very similar results, which means that those two matrices have similar properties. Our future work is to try to understand the inconsistencies between these two methods since both are linear models and should be yielding the same transition matrix A.

Confirmation of the DMDc Model. The neural network can also be used to test the DMDc model. We only need to increase the input dimension by the number of controls, and regard them as additional input. Also, as the transformation matrices are not square matrices, we only try to compare the singular values of the whole transformation matrices and the eigenvalues of the A matrices. Using the gender control as an example, the mean squared error (MSE) and comparison of singular value of DMDc and neural network with control is shown in figure 15. The same train-test split ratio is used as in the previous section for each of the 100 trials. From table 8, we see that DMDc has a slightly better performance than neural network. Also, the two matrices are still similar with each other.

Distinguishing Dropout Participants. According to the information from GRYD, after taking one or more questionnaires, some children may quit the program by various reasons. For example, they may be reluctant to join the program, or they may move during the program. We call those children the “dropout" group. There are two main reasons for participants to dropout of the program according to the GRYD team: long-term non-attendance and refusing services.

Among all 22548 participants of the program, 13549 of them has a record of their status. For these participants with a recorded status, 5258 students dropped the program before they are qualified to graduate. For those dropped out participants, 4110 of them only took the intake questionnaire before they dropout, 922 students took two questionnaires before dropping out, and the rest of the 226 participants took 3-5 questionnaires. On the other side, 4264 students graduated from the program successfully, 617 of them only took one questionnaire, 2256 of them took two questionnaires and the rest of the 1391 participants

	DMDc		NNc	
Trial	MSE (T)	MSE (W)	MSE (T)	MSE (W)
1	0.0987	0.0956	0.1004	0.0962
2	0.1013	0.0957	0.1023	0.0963
3	0.0996	0.0956	0.1008	0.0965
4	0.0984	0.0956	0.1015	0.0972
5	0.0980	0.0956	0.1016	0.0961
Mean	0.1010	0.0956	0.1016	0.0968
std	0.0018	0.00046	0.0020	0.00082

Table 8. Comparison of MSE of DMD with control and linear neural network with control. The result is formulated in the same way as table 7. The mean and standard deviation is calculated based on 100 trials. We can observe that the two approaches provide us with similar result, and the result is consistent with the theoretical lowest error bound discussed in the supplemental material.

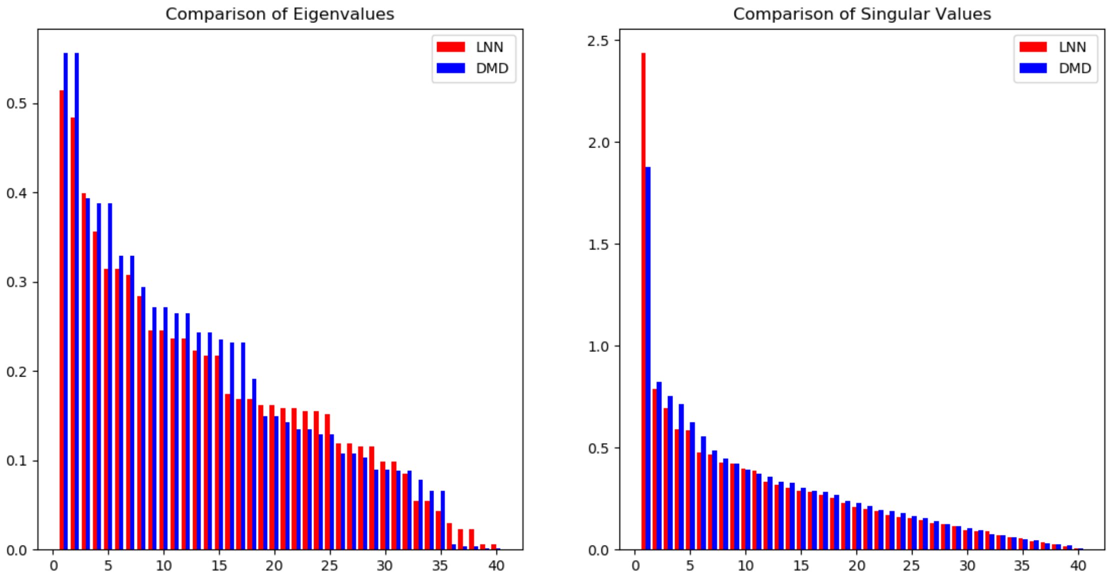

Fig. 15. Comparison of eigenvalues and singular values of DMD with control and Neural Network with Control. As the transformation matrices are not squared matrices. To compare the eigenvalues, we take the A matrix of DMDc (that is , the matrix not multiplying with the control variables) and the corresponding "A matrix" of linear neural network with control. We calculate the singular values of a matrix formed by vertically concatenating A and B. The result still shows that the eigenvalues and singular values derived from those matrices have high correlation coefficient. Therefore, the two matrices have similar properties.

took 3 or more questionnaires to graduate. For the remainder of the participants, 1686 of them are still in progress that are classified as “other" (including transferring to a different program).

We use machine learning techniques such as SVM and random forest (combination of decision trees) to see if the "dropout" group and "non-dropout" group are distinguishable. Using SVM as a classifier, we pick those who drop out after taking Y1 and label them with 0, and randomly choose the same number of children that didn’t drop out after taking Y1 and label them with 1. After trained by 80% of the dataset, SVM shows that it has 85.89% accuracy on the test set. Similarly, the random forest gives an accuracy of 89.42% on the test set. High performance on the test set means that no over-fitting occurs. Therefore, those models can be generalized to new data, which further indicates that the groups of "dropout" and "non-dropout" are separable by some classifiers. From the result of the classifiers, we may be able to find out which question plays an important rule. Also, it may be possible to design an approach to describe the probability of each participant dropping out from the program when they take the first questionnaire.

Other Future Works. In addition to the questions discussed above, there are additional questions we are interested in exploring. First of all, we want to figure out an appropriate way to produce a generalized matrix whose application is no longer limited to specific program period or a demographic characteristic. In other words, this generalized matrix can perform well on any kind of data. We can try modifying the existing algorithms to compute the generalized matrix by making the best use of all available information. Secondly, we want to look further into the data set to see if there is additional information that could be analyzed. In the future, we would like to extend our methods to include all questions, such as the behavioral and subjective questions. Furthermore, as we produced continuous variables for participants’ responses which are designed by questionnaires to be discrete, we can either refine our DMD methods or the format of our data sets to achieve higher prediction accuracy. Finally, we can analyze results produced by DMDc more carefully. In our case, we only compared the prediction accuracy between DMDc and DMD. However, we didn't spend much time on interpreting the transformation matrices, especially for the one before the control factor. We hope to understand better how the GRYD program have different effects on different groups of people.

ACKNOWLEDGMENTS. This work was sponsored by NSF grant DMS-1659676 and the City of Los Angeles Gang Reduction and Youth Development (GRYD) Analysis Program. We acknowledge Professor P. Jeffrey Brantingham for providing valuable information on GRYD program and Dr. Hanqin Cai for fruitful talks on algorithms for our analysis.

1. K M.Hennigan, KA Kolnick, F Vindel, CL Maxson, Targeting youth at risk for gang involvement: Validation of a gang risk assessment to support individualized secondary prevention. *Child Youth Serv. Rev.* **56**, 86–96 (2015).
2. JL Proctor, SL Brunton, J Kutz, Dynamic mode decomposition with control. *Siam J. Appl. Dyn. Syst.* **15**, 142–161 (2016).
3. DMLSLB Jonathan H. Tu, Clarence W. Rowley, JN Kutz, On dynamic mode decomposition: Theory and applications. *Am. Inst. Math. Sci.* **1**, 391–421 (2014).
4. SL Brunton, BW Brunton, JL Proctor, JN Kutz, Koopam invariant subspaces and finite linear representations of nonlinear dynamical systems for control. *PLoS ONE* (2016).
5. JL Proctor, SL Brunton, J Kutz, Dynamic mode decomposition with control. *Siam J. Appl. Dyn. Syst.* **15**, 142–161 (2016).
6. P Dangeti, Support vector machines and neural networks in *Statistics for Machine Learning.* (2018).
7. P Dangeti, Tree-based machine learning models in *Statistics for Machine Learning.* (2018).
8. TA Brown, The common factor model and explanatory factor analysis. (2015).
9. CW Rowley, I Mezić, S Bagheri, P Schlatter, DS Henningson, Spectral analysis of nonlinear flows. *J. Fluid Mech.* **64**, 115–127 (2009).
10. C Seger, An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing. *Degree Proj. KTH Royal Inst. Technol. Sch. Electr. Eng. Comput. Sci.* (2018).
11. M Feder, Time series analysis of repeated surveys: the state - space approach. *Stat. Neerlandica* **55**, 182–199 (2001).
12. M Raissi, P Perdikaris, GE Karniadakis, Multistep neural networks for data-driven discovery of nonlinear dynamical systems. *arXiv:1801.01236 [math.DS]* (2018).
13. CAW Paul P. Biemer, Sharon L. Christ, A general approach for estimating scale score reliability for panel survey data. *Psychol. Methods* **14**, 400–412 (2009).