

**MATCHING STUDENTS WITH SUPPORT SERVICES THROUGH A
CONSTRAINED LINEAR OPTIMIZATION MODEL**

A THESIS

Presented to the Department of Mathematics and Statistics
California State University, Long Beach

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in Mathematics

Committee Members:

Jen-Mei Chang, Ph.D. (Chair)
Chung-Min Lee, Ph.D.
Tianni Zhou, Ph.D.

College Designee:

Tangan Gao, Ph.D.

By Diana Gonzalez

B.S., 2015, California State University, Long Beach

August 2018

ABSTRACT

**MATCHING STUDENTS WITH SUPPORT SERVICES THROUGH A
CONSTRAINED LINEAR OPTIMIZATION MODEL**

By

Diana Gonzalez

August 2018

College students have several on-campus resources available to increase academic success and generally do not have an effective way for selecting the right one. We were inspired to create a decision aid to help students pick their resources due to the success of decision aids used in the medical field. An instrument to match students with the most appropriate support service by studying existing data was constructed by analyzing roughly 350 student responses to questions from the Ruffalo Noel Levitz survey and a questionnaire, totaling 108 questions, both administered by TRiO Student Support Programs at CSULB. One of the goals of this project is to make a shorter survey that would capture the same information as the original.

To that end, Principal Component Analysis was performed on the existing data to determine the number of categories needed for the survey. Factor Analysis was then used to select the representative question from each category. The new survey remains valid since the Ruffalo Noel Levitz survey was validated in the first place.

To develop our model for matching students with the most appropriate support services, we asked students in the Early Start Mathematics Program at California State University, Long Beach to answer the new survey and rank their preferred resources; we collected 300 data points, 162 of which are usable. The ranked resources were used as ground truth to create a training set. A constrained least-squares optimization model was proposed to match the survey responses with the ranked resources. The solution to the least-squares problem, solved via the method of Lagrangian-Multipliers, gave us a way to match survey responses

to resources. k -Fold Cross Validation was used to validate the accuracy of the method by using the ground truth gathered from the ranked resources.

ACKNOWLEDGEMENTS

I would like to begin by thanking Charity Bowles for providing the student responses to the Ruffalo Noel Levitz surveys. I would also like to thank Dr. Misty Sawatzky for the guidance in creating and validating the survey in our decision aid instrument. Thank you both for making this project possible.

I am grateful for the support of my thesis committee. Thank you Dr. Lee and Dr. Zhou for being encouraging and supportive. Dr. Lee, thank you for helping me use precise language and understand concepts. Dr. Zhou, thank you for reviewing the concept of Factor Analysis with me several times and for discussing the procedure for creating the survey. I value the time and patience you gave me. The motivational emails and pep talks helped me keep pushing forward.

To my friends, thank you for all of your support and motivation. Specifically, thank you Debbie for going to all of my talks, helping me form coherent sentences, and being there for me through this process. I appreciated the late Target runs and your help with Toby. Alberto, thank you for working with me late into the nights, for listening to me vent, and for being caring. Your presence this past year helped me finish my thesis with happier memories. To Kayla, thank you for not talking to me about my thesis and instead providing comical stories. To Josh, thank you for reminding me that line dancing is a good distraction from math.

I would like to thank my wonderful family for being understanding of my absence and for all the love they kept giving me. Thank you for showing me love through food, massages, and cariñitos. Mami y Papi, los quiero mucho; los amo con todo mi corazón. Muchas gracias por todo el amor y el apoyo que me han dado. Son los mejores padres del mundo! Brian, thank you for your availability and for listening when I needed to let my thoughts out. You're the best brother anyone could ask for.

Lastly, but most importantly, I would like to thank Dr. Jen-Mei Chang. Dr. Chang,

thank you for helping me believe in myself and for reading every abstract, presentation, paper, and the many submissions of this thesis. You have helped me grow as a mathematician and as a person. I am a better student, a more educated instructor, and a stronger person because of your guidance. I value all the time and patience you gave me over the last seven years. Thank you for working with me in creating a project I am proud of.

This research project was approved under the local Institutional Research Board (IRB) Reference Number 17-374 and Reference Number 17-347. This project was funded by the Office of Research and Sponsored Programs (ORSP) Summer Student Research Assistantship for the Summer of 2017 and the CSULB Graduate Research Fellowship for the 2017-2018 Academic Year.

TABLE OF CONTENTS

ABSTRACT	ii
ACKNOWLEDGEMENTS	iv
LIST OF TABLES	vii
LIST OF FIGURES	viii
1. INTRODUCTION	1
2. BACKGROUND LITERATURE	3
3. METHODS	5
4. EXPERIMENTAL RESULTS	24
5. DISCUSSION AND FUTURE WORKS	29
6. SUMMARY AND CONCLUSIONS	34
APPENDICES	36
A. RNL FORMS AND CSI SURVEYS	37
A RNL FORMS AND CSI SURVEYS	37
B. MATLAB CODES	46
B MATLAB CODES	46
REFERENCES	54

LIST OF TABLES

1	The Loadings of the Exams for the Two Factors	18
2	The Specific Variances of the Exams When Using Two Factors.....	18
3	Percent of Students That Selected Each Resource	30
4	Condition Numbers of Each Resource	35

LIST OF FIGURES

1	The plane $z = 2x - y$ with 10% perturbation.	12
2	The principal components and the data.....	13
3	The z -plane with 20% perturbation and its principal components.	14
4	The z -plane with 55% perturbation and its principal components.	15
5	A Visual Presentation of the k -Fold Cross Validation method.	23
6	The Singular Values of the ESM data gathered.	25

CHAPTER 1

INTRODUCTION

The need to change the way students learn has increased due to changes in student population as well as their environment. New resources are being implemented to address such changes. For example, cell phones are an accessory that most students have; there exist tools that can incorporate the use of cell phones in the classroom to enhance the learning experience. These resources include the use of technology in the classroom and tailoring the curriculum to individual learning [1].

Liberal Education and America's Promise (LEAP) is a program that addresses the need for higher levels of learning and knowledge for college students. It focuses on helping teachers help students by having students gain intellectual and practical skills that include, e.g., inquiry and analysis, quantitative literacy, teamwork and problem solving, and integrative and applied learning [2]. LEAP also recommends high-impact practices such as learning communities, undergraduate research, internships, etc. [2].

An observation to make about these resources is that instructors choose what resources to implement and in what way [1]. It is recommended that instructors implement different styles of teaching such as flipped classrooms, to help students take ownership of their learning. Instructors and advisors also present students with information regarding resources on campus that target the general population such as tutoring or attending office hours. Instructors and advisors are helping increase student success by selecting a certain teaching style to use or encourage students to partake in specific resources. However, students are receiving recommendations tailored to the general public instead of the individual. Students are less likely to follow through with such recommendations because recommendations may not pertain to them.

A similar phenomenon is often observed in medical settings when patients are asked to make decisions. A proposed solution to help patients make decisions is to incorporate the

use of decision aids. A decision aid is a tool in the form of a pamphlet, video, or poster, that provides information regarding a specific decision. For example, Mayo Clinic provided a decision aid titled The Statin Choice that helped their patients make an informed decision about taking statins as a treatment to reduce the risk of diabetes mellitus. The Statin Choice included information regarding the side effects of the medicine and other alternatives. When information on a topic is insufficient, a person's decision will fall into the hands of someone who knows more about the topic. The decision aid served in returning the ability to decide back to the patients [3].

In this project, we proposed a tool to match students with the support resources that are most suitable to their perceived needs. A decision aid was developed in the form of a survey and a list of resources tailored to our specific survey population. We proposed a method to match the survey responses to the list of resources to create a prediction model through a constrained optimization problem. The method of k -Fold Cross Validation was then used to determine the accuracy of the model.

The rest of this thesis is organized as follows. Chapter 2 discusses current techniques used to improve student success rates along with their strengths and weaknesses. Chapter 3 begins by introducing the techniques to analyze the data gathered from surveys. Then it looks into data reduction techniques used to create the decision aid. Lastly, the proposed model used to determine recommendations to students is introduced along with the method used to validate the model. The finalized decision aid and its strengths and weaknesses are analyzed in Chapter 4. Chapter 5 discusses future works and methods of improvement for the model. The study ends with Chapter 6, a summary of the work and conclusion.

CHAPTER 2

BACKGROUND LITERATURE

The transition from high school to college introduces many factors that may add stress on students such as the adjustment of paying for their education, longer commutes, changing locations, acquiring new friendships, learning how to study, and many more. Freshmen are typically given various resources to aid in their academic success and overcome these stressors. Examples of resources include information about financial aid, living communities on campus, and tutoring. Stressors, habits, and experiences contribute to the decisions students will make in regards to their academic well being and personal life style choices [4].

There are also resources for instructors and administrators that identify *at-risk* factors in students. At-risk factors are the characteristics that identify students that may not have the tools needed to succeed academically. It is critical to identify students who may have such factors early in their academic career to increase their chances of success [5].

An example of a tool used by instructors to identify at-risk factors is a dropout model. A specific type of dropout model implemented by Duarte et al. [6] approaches the problem of student dropout by analyzing external factors such as familial support, institutional integration, and self-efficacy. The model analyzed student academic data and administrative records to identify at-risk students.

Similarly, there exist surveys that are used for the same purpose of identifying at-risk students and their behaviors at academic institutions. The information gathered is intended to be studied to make improvements in a program or at an institution. Often times, the results of these surveys are left unstudied due to large amounts of data, not knowing what is expected, etc.

To summarize, incoming freshmen are beginning a new phase in their academic careers that requires students to adapt to a new environment and acquire new skills and responsibilities. Such changes may prevent students from succeeding academically, therefore tools are

being introduced to academic advisors and instructors to identify students that are at-risk of not succeeding. Measuring the success of those tools can be difficult due to the way data gathered therefore data is left unstudied, preventing improvements from being made.

For this study we analyzed unstudied data from surveys containing questions from the Ruffalo Noel Levitz (RNL) Forms. The surveys are used to identify incoming freshmen that may be considered to be at-risk by addressing *student's academic motivations, areas of risk*, and *receptivity to specific student services*. This type of survey is meant to help advisors or faculty intervene with students earlier in their academic career to provide students with the necessary resources [7]. We want to develop an instrument that identifies at-risk characteristics in students and introduces the student to support services that address their specific needs.

CHAPTER 3

METHODS

Data Normalization

Data attained from surveys can be difficult to analyze without some type of cleaning or pre-processing because responses can come in different forms depending on the type of questions. Answers from free response questions are recorded as text characters compared to responses from multiple choice and Likert scale (Strongly Agree to Strongly Disagree) questions which are recorded as numbers corresponding to the placement of the response.

The current study is concerned with multiple choice and Likert scale questions. Individual responses are realized as rows in a spreadsheet where each column stores a numerical response to a particular survey question for the same participant. A matrix representation for a P -participant- N -question survey is given by a P -by- N matrix

$$X = \begin{matrix} & Q_1 & Q_2 & Q_3 & \cdots & Q_N \\ \begin{matrix} P_1 \\ P_2 \\ \vdots \\ P_P \end{matrix} & \begin{bmatrix} 1 & 4 & 0 & \cdots & 5 \\ 3 & 6 & 1 & \cdots & 4 \\ \vdots & & & \ddots & \vdots \\ 2 & 4 & 2 & \cdots & 7 \end{bmatrix} \end{matrix},$$

where the $(i, j)^{th}$ element in X represents i^{th} participant's level's agreement to the j^{th} question in the survey.

A common first step in survey data analysis is to transform the data into a uniform response scale, which can be done via the map $\phi : R \rightarrow [0, 1]$,

$R = \{1, 2, \dots, k\}$ is the set of response values. The map is defined by $\phi(i) = \frac{i-1}{k-1}$, where i is the selected response and k is the number of possible responses for a specific question. For example, consider responses of participant one to questions one and two. Suppose question

one had three possible responses and question two had five possible responses. His/her response to question one, $i = 2$, will be converted to $\phi(2) = \frac{1}{2}$ with $k = 3$, and the response to question two, $i = 4$, will be converted to $\phi(4) = \frac{3}{4}$ with $k = 5$. Thus, normalization gives

$$P_1 \begin{matrix} & Q_1 & Q_2 \\ \begin{bmatrix} 1 & 4 \end{bmatrix} \end{matrix} \Rightarrow P_1 \begin{matrix} & Q_1 & Q_2 \\ \begin{bmatrix} \frac{1}{2} & \frac{3}{4} \end{bmatrix} \end{matrix}$$

Moreover, results from surveys often correspond to responses with different units or meanings. Standardizing the data and analyzing their *z-scores* allow us to compare responses. From now on, the data matrix X will correspond to the standardized responses fairly. Implementation of the techniques discussed can be found in Appendix B.1 and B.3.

Data Reduction

Data redundancy is very common in surveys since many questionnaires are designed to have multiple questions assessing similar categories of information for the purpose of validation. *Principal Component Analysis* (PCA) is a technique that can effectively reveal uncorrelated variables to represent the original data set using a lot less information, allowing us to reduce the number of questions in a survey without losing much information.

PCA proceeds by calculating the eigenvectors of a covariance matrix also known as the *principal components*. The derivation of PCA, which we review next, follows closely with the presentation given in [8].

PCA Derivation

Suppose there exists a set of points $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(p)}\}$ such that $\mathbf{x}^{(i)} \in \mathbb{R}^n$ for large n and $i = 1, \dots, p$. We want to find an appropriate map that projects data points in their ambient space, which is typically high-dimensional, to a much lower dimensional space while retaining as much variance as possible. In other words, we look for a map $f : \mathbb{R}^n \rightarrow \mathbb{R}^l$ where $l \ll n$.

Suppose a map, $g : \mathbb{R}^l \rightarrow \mathbb{R}^n$, exists such that $g(f(\mathbf{x})) = D \cdot f(\mathbf{x})$ acts as the inverse projection of f , cast as a matrix multiplication problem. The goal of PCA is to obtain a matrix, D , where DD^T is an orthogonal projection matrix and $g(f(\mathbf{x})) = \tilde{\mathbf{x}}$, is as close to \mathbf{x} as possible. Equivalently, PCA finds the optimal D that minimizes the objective function $J(f(\mathbf{x})) = \|\mathbf{x} - \tilde{\mathbf{x}}\|_2^2$. Here, let $\mathbf{c} = f(\mathbf{x})$ for the ease of notation.

To find D , we begin by first examining the following optimization problem:

$$\min J(\mathbf{c}) = \min \|\mathbf{x} - \tilde{\mathbf{x}}\|_2^2 = \min \|\mathbf{x} - g(f(\mathbf{x}))\|_2^2 = \min \|\mathbf{x} - g(\mathbf{c})\|_2^2 \quad (3.1)$$

We begin to solve the optimization problem by expanding the objective function using the definition of the 2-norm:

$$\begin{aligned} & \arg \min_{\mathbf{c} \in \mathbb{R}^l} (\mathbf{x} - g(\mathbf{c}))^T (\mathbf{x} - g(\mathbf{c})) \\ &= \arg \min_{\mathbf{c} \in \mathbb{R}^l} \mathbf{x}^T \mathbf{x} + g(\mathbf{c})^T g(\mathbf{c}) - \mathbf{x}^T g(\mathbf{c}) - g(\mathbf{c})^T \mathbf{x} \\ &= \arg \min_{\mathbf{c} \in \mathbb{R}^l} g(\mathbf{c})^T g(\mathbf{c}) - 2g(\mathbf{c})^T \mathbf{x}. \end{aligned}$$

Notice the term $\mathbf{x}^T \mathbf{x}$ is a constant, free of \mathbf{c} , therefore it can be disregarded. Since $g(\mathbf{c}) = D \cdot \mathbf{c}$, we get the following equation:

$$\begin{aligned} & \arg \min_{\mathbf{c} \in \mathbb{R}^l} \mathbf{c}^T D^T D \mathbf{c} - 2\mathbf{c}^T D^T \mathbf{x} \\ &= \arg \min_{\mathbf{c} \in \mathbb{R}^l} \mathbf{c}^T I_l \mathbf{c} - 2\mathbf{c}^T D^T \mathbf{x} \quad (\text{since } D \text{ is orthogonal}) \\ &= \arg \min_{\mathbf{c} \in \mathbb{R}^l} \mathbf{c}^T \mathbf{c} - 2\mathbf{c}^T D^T \mathbf{x}. \end{aligned}$$

$\nabla_{\mathbf{c}} J = 0$ is a necessary condition for the global minimum of the optimization problem in Equation (3.1). Notice that $\nabla_{\mathbf{c}} J = 2\mathbf{c} - 2D^T \mathbf{x} = 0$. Thus $\mathbf{c} = D^T \mathbf{x}$ and by substitution, $f(\mathbf{x}) = D^T \mathbf{x}$. Hence, $g(f(\mathbf{x})) = DD^T \mathbf{x}$, where DD^T is an orthogonal projection matrix

projecting the point from a high dimensional space, \mathbb{R}^n , to a low dimensional space, \mathbb{R}^l .

This is true for a single point, \mathbf{x} .

Now, let $X = \begin{bmatrix} -\mathbf{x}^{(1)} - \\ \vdots \\ -\mathbf{x}^{(p)} - \end{bmatrix}$ be the data ensemble matrix whose row vectors are distinct

data points in the set, $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(p)}\}$, such that $X \in \mathbb{R}^{p \times n}$. Recasting Equation (3.1) for the entire data ensemble gives us

$$\min \sum_{i=1}^p \|\tilde{\mathbf{x}}^{(i)} - \mathbf{x}^{(i)}\|_2^2 = \min_{D \in \mathbb{R}^{n \times l}} \sum_{i=1}^p \|DD^T \mathbf{x}^{(i)} - \mathbf{x}^{(i)}\|_2^2.$$

Consider the case $l = 1$ where $D = \mathbf{d}^{(1)}$ is a column vector. Then the problem can be written in terms of matrices: $\min_{\mathbf{d} \in \mathbb{R}^n} \|(X - \mathbf{d}\mathbf{d}^T X)^T\|_F^2 = \min_{\mathbf{d} \in \mathbb{R}^n} \|X^T - X^T \mathbf{d}\mathbf{d}^T\|_F^2$ subject to $\mathbf{d}^T \mathbf{d} = 1$. The objective function can be expanded similarly to the one previously done; let $X^T = \tilde{X}$:

$$\begin{aligned} & \arg \min_{\mathbf{d} \in \mathbb{R}^n} \text{Tr}((\tilde{X} - \tilde{X}\mathbf{d}\mathbf{d}^T)^T(\tilde{X} - \tilde{X}\mathbf{d}\mathbf{d}^T)) \\ = & \arg \min_{\mathbf{d} \in \mathbb{R}^n} \text{Tr}(\tilde{X}^T \tilde{X} + \mathbf{d}\mathbf{d}^T \tilde{X}^T \tilde{X} \mathbf{d}\mathbf{d}^T - \mathbf{d}\mathbf{d}^T \tilde{X}^T \tilde{X} - \tilde{X}^T \tilde{X} \mathbf{d}\mathbf{d}^T) \\ = & \arg \min_{\mathbf{d} \in \mathbb{R}^n} \text{Tr}(\tilde{X}^T \tilde{X}) + \text{Tr}(\mathbf{d}\mathbf{d}^T \tilde{X}^T \tilde{X} \mathbf{d}\mathbf{d}^T) - 2\text{Tr}(\mathbf{d}\mathbf{d}^T \tilde{X}^T \tilde{X}) \\ = & \arg \min_{\mathbf{d} \in \mathbb{R}^n} \text{Tr}(\mathbf{d}\mathbf{d}^T \tilde{X}^T \tilde{X} \mathbf{d}\mathbf{d}^T) - 2\text{Tr}(\mathbf{d}\mathbf{d}^T \tilde{X}^T \tilde{X}) \end{aligned}$$

The term $\text{Tr}(\tilde{X}^T \tilde{X})$ is free of D therefore it can be disregarded. The order of square matrices in a trace, $\mathbf{d}\mathbf{d}^T$ and $\tilde{X}^T \tilde{X}$, can be rearranged to use the property $\mathbf{d}^T \mathbf{d} = 1$ to simplify the

problem.

$$\begin{aligned}
& \arg \min_{\mathbf{d} \in \mathbb{R}^n} \quad \text{Tr}(\tilde{X}^T \tilde{X} \mathbf{d} \mathbf{d}^T \mathbf{d} \mathbf{d}^T) - 2\text{Tr}(\tilde{X}^T \tilde{X} \mathbf{d} \mathbf{d}^T) \\
&= \arg \min_{\mathbf{d} \in \mathbb{R}^n} \quad \text{Tr}(\tilde{X}^T \tilde{X} \mathbf{d} \mathbf{d}^T) - 2\text{Tr}(\tilde{X}^T \tilde{X} \mathbf{d} \mathbf{d}^T) \text{ since } \mathbf{d}^T \mathbf{d} = 1 \\
&= \arg \min_{\mathbf{d} \in \mathbb{R}^n} \quad -\text{Tr}(\tilde{X}^T \tilde{X} \mathbf{d} \mathbf{d}^T) \\
&= \arg \min_{\mathbf{d} \in \mathbb{R}^n} \quad -\text{Tr}(\mathbf{d}^T \tilde{X}^T \tilde{X} \mathbf{d})
\end{aligned}$$

The equivalent optimization problem is $\max_{\mathbf{d} \in \mathbb{R}^n} \text{Tr}(\mathbf{d}^T \tilde{X}^T \tilde{X} \mathbf{d})$ subject to $\mathbf{d}^T \mathbf{d} = 1$. Now, let $A = \tilde{X}^T \tilde{X}$, where A is real symmetric matrix. Since the argument of the trace, $\mathbf{d}^T A \mathbf{d}$, is a scalar, $\text{Tr}(\mathbf{d}^T A \mathbf{d}) = \mathbf{d}^T A \mathbf{d}$. Let $\lambda = \mathbf{d}^T A \mathbf{d}$ such that (λ, \mathbf{d}) is an eigenpair of matrix A . Then

$$\begin{aligned}
A \mathbf{d} &= \lambda \mathbf{d} \\
(A \mathbf{d})^T &= (\lambda \mathbf{d})^T \\
\mathbf{d}^T A^T &= \lambda \mathbf{d}^T \\
\mathbf{d}^T A &= \lambda \mathbf{d}^T \quad A \text{ is symmetric} \\
\mathbf{d}^T A \cdot \mathbf{d} &= \lambda \mathbf{d}^T \cdot \mathbf{d} \\
\mathbf{d}^T A \cdot \mathbf{d} &= \lambda.
\end{aligned}$$

This optimization problem of $\max_{\mathbf{d} \in \mathbb{R}^n} \lambda$, becomes the eigenvalue problem

$$\tilde{X}^T \tilde{X} \mathbf{d}^{(1)} = \lambda_1 \mathbf{d}^{(1)}$$

where $\mathbf{d}^{(1)}$ is the eigenvector corresponding to the largest eigenvalue. Similarly, $\tilde{X}^T \tilde{X} \mathbf{d}^{(i)} = \lambda_i \mathbf{d}^{(i)}$ where $\mathbf{d}^{(i)}$ is the i^{th} eigenvector of $\tilde{X}^T \tilde{X}$ with corresponding eigenvalue, λ_i . For the case proved above, $i = 1$, the first principal component, $\mathbf{d}^{(1)}$, associated with the largest

eigenvalue, identifies the direction with the most variability within the data. In contrast, the last principal component, $\mathbf{d}^{(l)}$, captures the least amount of variability [9]. We have shown that finding the optimal basis, $\{\mathbf{d}^{(1)}, \dots, \mathbf{d}^{(l)}\}$, to represent the empirical data stored in the columns of \tilde{X} , is equivalent to finding the eigenvectors of $\tilde{X}^T \tilde{X}$.

Singular Value Decomposition

We can determine the number of principal components needed to reduce the data while keeping as much statistical variance as possible by examining the singular values. The SVD Theorem guarantees a factorization of the form $U\Sigma V^T$ for any m -by- n matrix, X , where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal and $\Sigma \in \mathbb{R}^{m \times n}$ is a diagonal matrix with decreasing singular values σ_i 's. It turns out that the principal component directions of the data precisely coincide with the column vectors of the U matrix. To see this, consider $X = U\Sigma V^T$, then $X^T = V\Sigma^T U^T$. We then have

$$\begin{aligned}
XX^T &= U\Sigma V^T V\Sigma^T U^T \quad r = \text{rank}(X) \\
&= U\Sigma\Sigma^T U^T \\
&= U \left[\begin{array}{ccc|c} \sigma_1^2 & & & 0 \\ & \ddots & & 0 \\ & & \sigma_r^2 & 0 \\ \hline 0 & \dots & & 0 \end{array} \right]_{m \times m} U^T.
\end{aligned}$$

Rewriting it gives us

$$XX^T U = U \left[\begin{array}{ccc|c} \sigma_1^2 & & & 0 \\ & \ddots & & 0 \\ & & \sigma_r^2 & 0 \\ \hline 0 & \dots & & 0 \end{array} \right].$$

In particular, each column of U satisfies the equation $XX^T u^{(i)} = \sigma_i^2 u^{(i)}$, $1 \leq i \leq r$. That is, the eigenvector of XX^T that is associated with the eigenvalue of σ_i^2 is precisely the i^{th} column vector of U .

In this study, the principal components correspond to questions from the CSI Survey. Since the number of participants, P , was greater than the number of questions, N , in the matrix $X \in \mathbb{R}^{N \times P}$, the singular values are obtained from the covariance matrix $XX^T \in \mathbb{R}^{N \times N}$ instead of $X^T X$. We use the covariance matrix XX^T because it reduces computational costs since it is of lower dimensionality. We chose to capture a specific amount of statistical variance to determine the number of independent variables needed from the CSI survey by analyzing the singular values. This method determines the number of questions needed in the new survey and can be seen in Appendix B.1.

We now illustrate the effectiveness of PCA in data reduction with a simple example. Here, we have 300 points sampled from the plane $z = 2x - y$, as shown in Figure 1. We manually perturbed 10% of the data to include some ‘noise’, illustrated by the asterisk points in Figure 1.

Figure 2 shows that 99.37% of the statistical variance is attained with the first two singular values, $\sigma_1 = 207.9733$ and $\sigma_2 = 49.9355$. The third singular value, $\sigma_3 = 17.0774$, captures 0.63% statistical variance. Thus, the dimensionality of the data is two and the optimal basis is given by first two principal components.

Figure 3 and Figure 4 illustrate the same example with added noise. In Figure 3, 20% of the data is perturbed and, like in the first example, most of the variance is captured

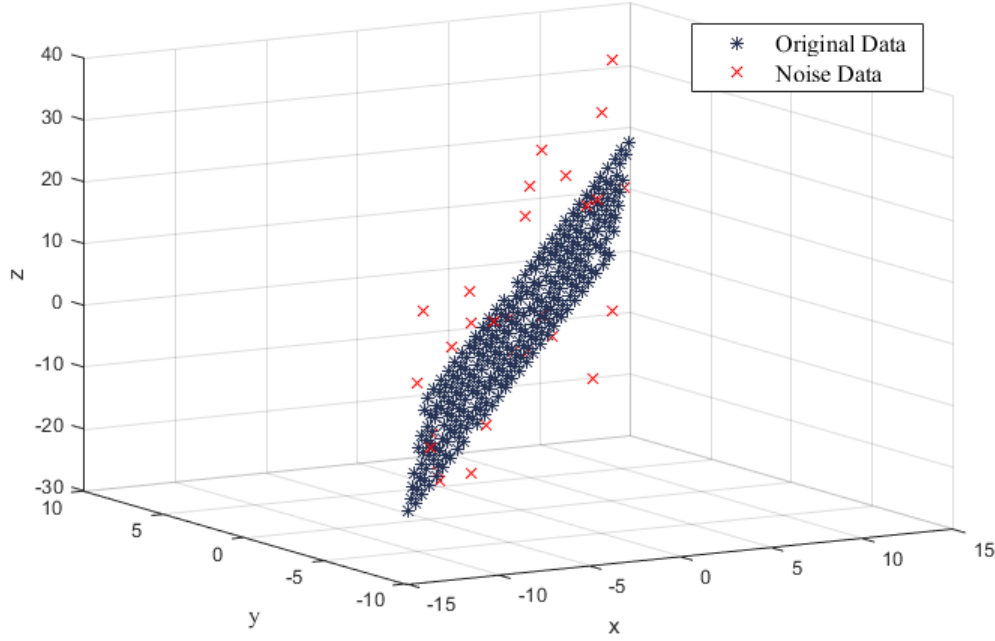


FIGURE 1. The plane $z = 2x - y$ with 10% perturbation.

using the first components. The first singular value, $\sigma_1 = 213.3792$, captures 92.46% of the statistical variance, the second singular value, $\sigma_2 = 58.1844$, captures 6.87% of the variance, and the third singular value, $\sigma_3 = 18.0693$, captures 0.66% of the variance.

Lastly, in Figure 4, 55% of the data was manually perturbed and still, the first singular value was significantly larger and the third singular value is close to zero. The singular values for this data are $\sigma_1 = 232.9518$, 78.3240, and 19.0989. They capture 89.30%, 10.10%, and 0.6% statistical variance, respectively. Notice that as we increase the amount of noise, the second singular value increases faster than the third singular value. So, we see that the dimensionality of the data is 2.

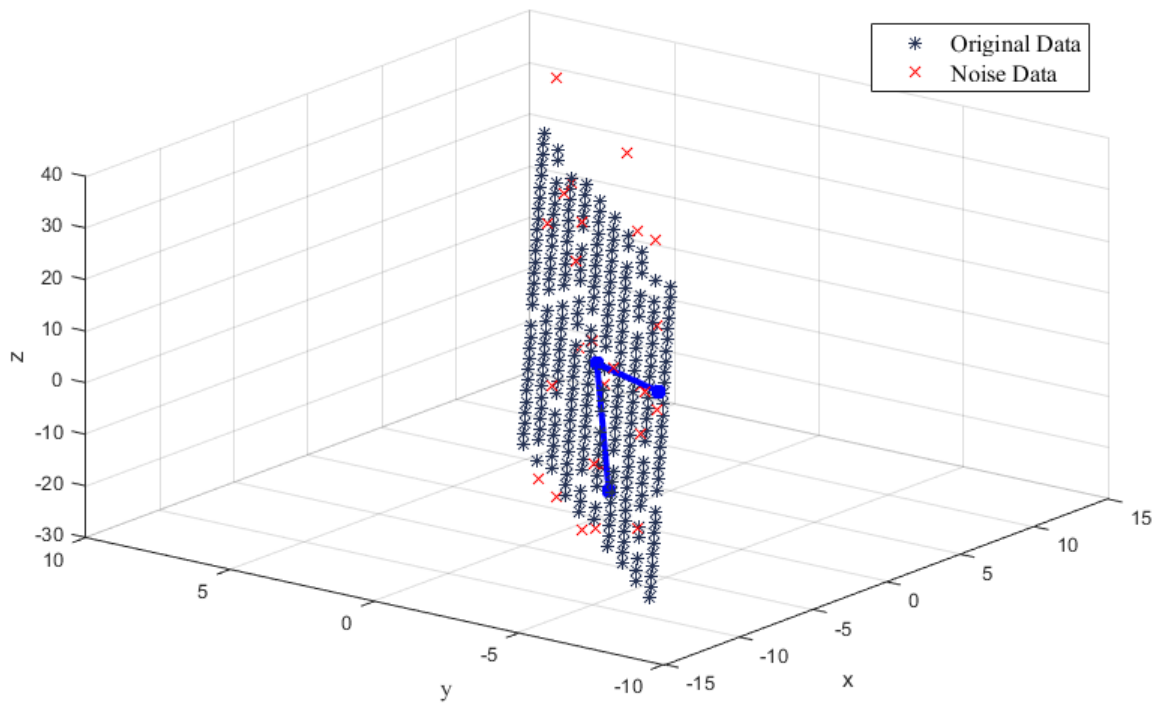


FIGURE 2. The principal components and the data.

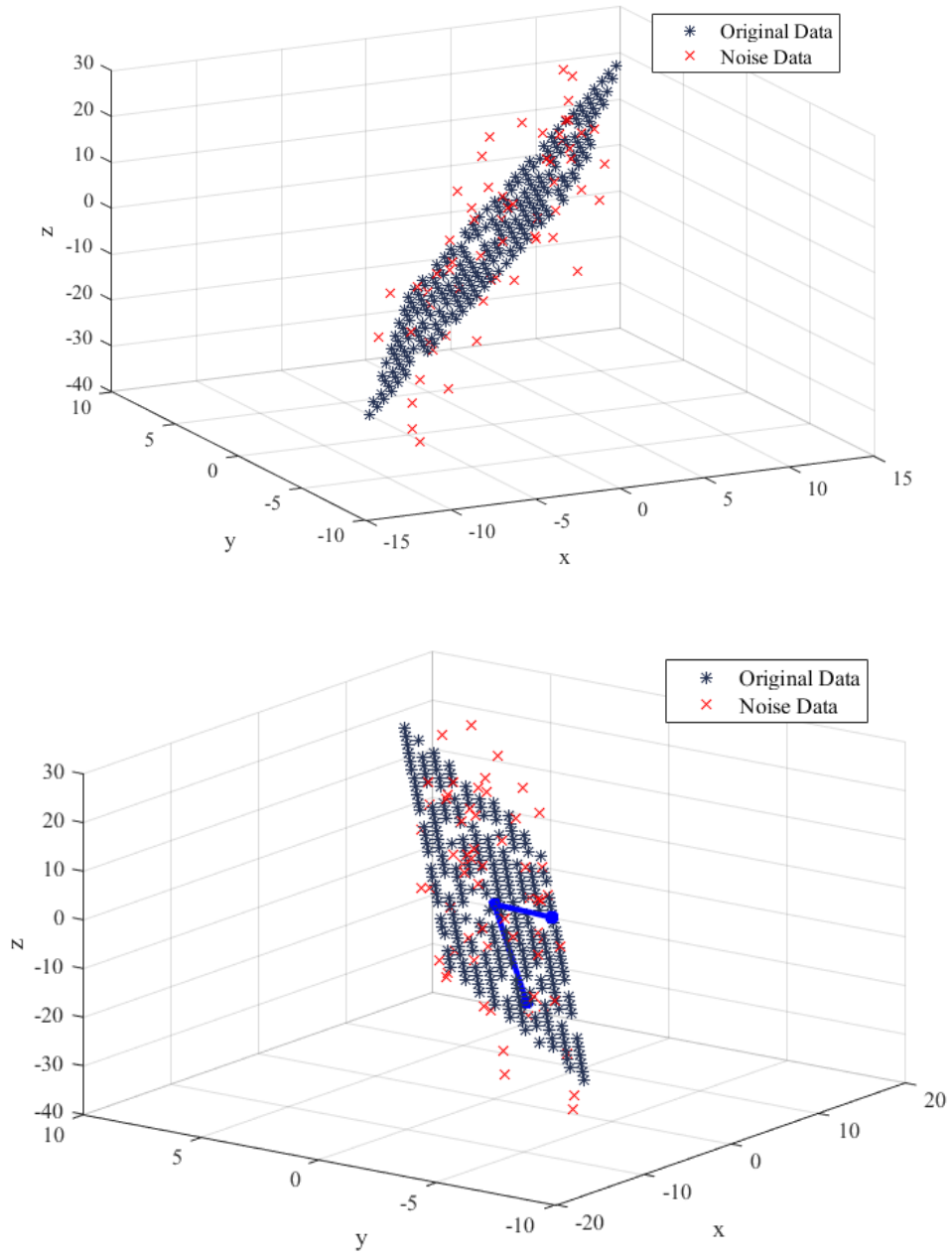


FIGURE 3. The z -plane with 20% perturbation and its principal components.

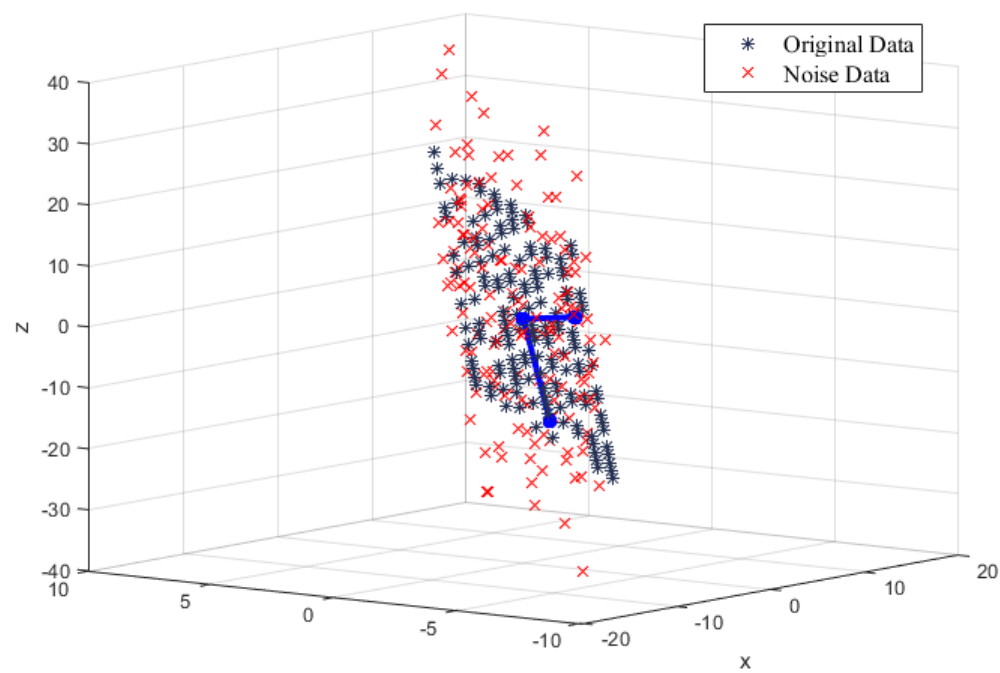
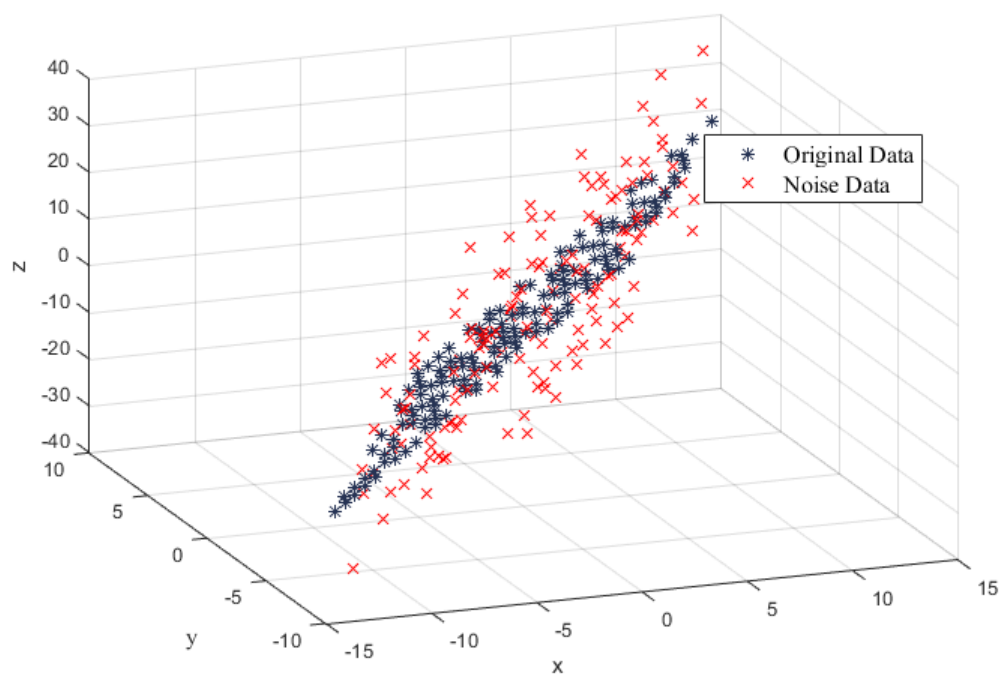


FIGURE 4. The z -plane with 55% perturbation and its principal components.

Data Refinement

Once the number of questions needed is determined, *Factor Analysis* (FA) is used to choose the appropriate questions for the new survey. Suppose q questions were asked to P participants; their responses are given by a P -by- q data matrix,

$$X = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_q \end{bmatrix}$$

where \mathbf{x}_j , $j = 1, 2, \dots, q$, are column vectors representing all participants' responses to the j^{th} question. FA reduces the dimensionality of the data matrix by assuming that there exists $N \leq q$ common factors, \mathbf{f}_l , that can be used to express the responses to the questions as linear combinations; that is, for each question in the survey, we can express the responses mathematically as follows:

$$\begin{aligned} \mathbf{x}_1 &= \lambda_{1,1}\mathbf{f}_1 + \lambda_{1,2}\mathbf{f}_2 + \cdots + \lambda_{1,N}\mathbf{f}_N + \eta_1 \\ \mathbf{x}_2 &= \lambda_{2,1}\mathbf{f}_1 + \lambda_{2,2}\mathbf{f}_2 + \cdots + \lambda_{2,N}\mathbf{f}_N + \eta_2 \\ &\vdots \\ \mathbf{x}_q &= \lambda_{q,1}\mathbf{f}_1 + \lambda_{q,2}\mathbf{f}_2 + \cdots + \lambda_{q,N}\mathbf{f}_N + \eta_q \end{aligned}$$

where $\lambda_{j,l}$, $j = 1, 2, \dots, q$; $l = 1, 2, \dots, N$ are the *factor loadings* or the weights for each factor, \mathbf{f}_l ; and η_j 's are the error terms corresponding to each question [9].

The model can be written in matrix form as such: $\mathbf{x} = \Lambda \mathbf{f} + \boldsymbol{\eta}$ where Λ is the matrix of factor loadings for an observation, \mathbf{x}_j , and factors, \mathbf{f}_l . The following assumptions are made in the model:

1. The factors, \mathbf{f}_l , are independently and identically distributed with mean 0 and variance

I

2. The error terms, η_j , are independently distributed with mean 0 and specific variance Ψ_j
3. The factors and errors are independent.

The amount of specific variance indicates how different the variable is from the factors. In other words, a high specific variance signifies that the corresponding variable is not well represented by the factors [10]. To determine the variables that best represent the data, we must look at factors with high loadings as well as low specific variances.

To find the matrix of factor loadings, Λ , we solve the following system of equations derived by the covariance of the model:

$$\begin{aligned}
\mathbf{x} &= \Lambda \mathbf{f} + \eta \\
\text{Cov}(\mathbf{x}) &= \text{Cov}(\Lambda \mathbf{f} + \eta) \\
&= \text{Cov}(\Lambda \mathbf{f}) + \text{Cov}(\eta) \quad \text{by Assumption (3)} \\
&= \Lambda \text{Cov}(\mathbf{f}) \Lambda^T + \text{Cov}(\eta) \\
&= \Lambda \Lambda' + \Psi \quad \text{by Assumptions (1) and (2)}.
\end{aligned}$$

As an example, suppose 100 students each took five exams. Two of the exams were on Mathematics, the other two on English, and the last one was a comprehensive exam. Intuitively, it would seem that there exists a correlation between the exam grades of the subject specific exams, in other words the two Mathematics exams are correlated and the two English exams correlated, and the comprehensive exam. FA can determine an appropriate way to factor the data.

To employ FA, the user selects the number of factors to categorize the data. Suppose we want to use two factors to represent the data. The loadings for the two factors are then used to determine what each factor represents. As seen in Table 1, Factor 1, \mathbf{f}_1 , has large

loadings for all of the exams which implies that \mathbf{f}_1 represents all of the exams evenly, thus one could say \mathbf{f}_1 represents *overall ability*. Factor 2, \mathbf{f}_2 ; on the other hand, has large loadings in the absolute value sense, for the Mathematics and English exams. Another observation is that the loadings of the Mathematics exams are of the opposite sign of the loadings from the English exams. One could say that \mathbf{f}_2 represents *subject specific abilities*.

The specific variance for data grouped by specific factors indicates how much each random

TABLE 1. The Loadings of the Exams for the Two Factors

Exam Type	Factor 1	Factor 2
Math Exam 1	0.7193	0.7256
Math Exam 2	0.6618	0.7189
English Exam 1	0.6784	-0.6090
English Exam 2	0.7649	-0.6170
Comprehensive Exam	0.8876	-0.2627

variable varies from the defined group. A high specific variance, that is, a variance close to 1, indicates that the corresponding variable cannot be represented in terms of the factors where a lower specific variance, a variance closer to 0, indicates that the random variable can be represented in terms of all the factors [10]. The estimated specific variances for the data using two factors are seen in Table 2. Notice that the Comprehensive Exam has a very low variance. This implies that the Comprehensive Exam is well represented by the two factors.

TABLE 2. The Specific Variances of the Exams When Using Two Factors

Exam Type	Specific Variance
Math Exam 1	0.3829
Math Exam 2	0.2031
English Exam 1	0.3512
English Exam 2	0.4321
Comprehensive Exam	0.1944

In this study, FA is used to derive a small subset of questions that will be used in the

classification model. Its implementation is found in Appendix B.1.

Classification Model

Constrained Optimization Problem

Students' ranked resources are connected to the RNL questions in terms of self-efficacy, perceived notions in confidence, and their response to student services. For example, we expect students who have a high financial need to rank a resource associated with financial guidance higher than others. We propose to use the following constrained optimization problem to match students with the appropriate resource:

$$\min_{\substack{\mathbf{c} \in \mathbb{R}^N \\ \mathbf{w} \in \mathbb{R}^k}} f(\mathbf{c}; \mathbf{w}) \text{ subject to } \mathbf{w}^T \mathbf{w} = 1 \quad (3.2)$$

where $f(\mathbf{c}; \mathbf{w}) = \|X\mathbf{c} - Y\mathbf{w}\|_2^2$, $X_{P \times N}$ is a matrix of the responses to the survey, $Y_{P \times k}$ is a matrix of participants' ranked resources, and \mathbf{c} and \mathbf{w} are the corresponding weight vectors. The structure of the objective function, f , for our problem can be visualized as follows:

$$f(\mathbf{c}; \mathbf{w}) = \left\| \underbrace{\begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,N} \\ x_{2,1} & x_{2,2} & \dots & x_{1,N} \\ \vdots & & \ddots & \vdots \\ x_{P,1} & x_{P,2} & \dots & x_{P,N} \end{bmatrix}}_X \underbrace{\begin{bmatrix} c_1 \\ \vdots \\ c_N \end{bmatrix}}_{\mathbf{c}} - \underbrace{\begin{bmatrix} y_{1,1} & y_{1,2} & \dots & y_{1,k} \\ y_{2,1} & y_{2,2} & \dots & y_{1,k} \\ \vdots & & \ddots & \vdots \\ y_{P,1} & y_{P,2} & \dots & y_{P,k} \end{bmatrix}}_Y \underbrace{\begin{bmatrix} w_1 \\ \vdots \\ w_k \end{bmatrix}}_{\mathbf{w}} \right\|_2^2$$

The constraint of $\|\mathbf{w}\|_2^2 = 1$ is imposed to ensure an analytic solution of \mathbf{w} exists. The problem amounts to solving for \mathbf{c} and \mathbf{w} simultaneously. To do this, we will use the method of Lagrange Multipliers. First, define the Lagrangian, $\mathcal{L}(\mathbf{c}, \mathbf{w}, \lambda) = f(\mathbf{c}, \mathbf{w}) - \lambda \cdot g(\mathbf{c}, \mathbf{w})$,

where $g(\mathbf{c}, \mathbf{w}) = \|\mathbf{w}\|_2^2 - 1$. Substitute in the definitions to obtain

$$\mathcal{L}(\mathbf{c}, \mathbf{w}, \lambda) = \mathbf{c}^T X^T X \mathbf{c} - 2 \mathbf{w}^T Y^T X \mathbf{c} + \mathbf{w}^T Y^T Y \mathbf{w} - \lambda(\mathbf{w}^T \mathbf{w} - 1).$$

The local minimizers of the optimization problem come from the solution to the following system of equations:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{c}} = X^T X \mathbf{c} - X^T Y \mathbf{w} = 0 \quad (3.3)$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = Y^T X \mathbf{c} - Y^T Y \mathbf{w} + \lambda \mathbf{w} = 0 \quad (3.4)$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = -\mathbf{w}^T \mathbf{w} + 1 = 0 \quad (3.5)$$

Write \mathbf{c} in terms of \mathbf{w} in Equation (3.3): $\mathbf{c} = (X^T X)^{-1} X^T Y \mathbf{w}$. By substituting \mathbf{c} into Equation (3.4), we get

$$(Y^T Y - \lambda I) \mathbf{w} = Y^T X (X^T X)^{-1} X^T Y \mathbf{w}. \quad (3.6)$$

We then get the eigenvalue problem, $(A - \lambda I) \mathbf{w} = B \mathbf{w}$ where $A = Y^T Y$ and $B = Y^T X (X^T X)^{-1} X^T Y$. We can solve for λ using Equation (3.6) for when $\mathbf{w} \neq 0$.

$$(Y^T Y - \lambda I) \mathbf{w} = Y^T X (X^T X)^{-1} X^T Y \mathbf{w}$$

$$(Y^T Y - \lambda I - Y^T X (X^T X)^{-1} X^T Y) \mathbf{w} = 0$$

$$(Y^T Y - \lambda I - Y^T X X^{-1} (X^T)^{-1} X^T Y) \mathbf{w} = 0$$

$$(Y^T Y - \lambda I - Y^T Y) \mathbf{w} = 0$$

$$\lambda I \mathbf{w} = 0$$

Since $\mathbf{w} \neq 0 \Rightarrow \lambda = 0$ thus we obtain the eigenvalue problem $(A - B)\mathbf{w} = 0$ which is used to solve for \mathbf{w} , as done in Appendix B.5. The derived \mathbf{w} is then used to obtain \mathbf{c} .

k -Fold Cross Validation

Due to the fact that only a small amount of data points are available, we chose to generate the classification statistics with a version of the k -Fold Cross Validation technique. Cross validation methods are commonly used to test the stability and accuracy of the models in classification algorithms. Such methods are also used to test the accuracy of a model on a training data set to allow for improvements, before applying the model real data [11].

In k -fold cross validation, the data is divided into k subgroups where $(k - 1)$ groups are used to train the model and one group is used to test it. The process is repeated k times to allow each group to be a test group, as shown in Appendix B.4.

For example, suppose we are given the responses to the reduced survey and the ranked resources of 270 participants, the data is first divided into $k = 9$ subgroups each consisting of 30 participants' responses. The validation process begins by training the model using the ground truth responses afforded by the 8 groups; or equivalently, 240 responses. Now, let

$$X_{train} = \begin{bmatrix} -\mathbf{x}^{(1)} - \\ \vdots \\ -\mathbf{x}^{(p)} - \end{bmatrix}, p \in \{1, \dots, 240\},$$

be the matrix of the participants' responses that will be used to train the model. X_{train} is then separated into 8 matrices where each matrix, X_r , $r = 1, \dots, 8$, contains responses of the participants who chose resource r as their most preferred resource. The responses, X_r and the corresponding resource rankings, Y_r , are used to solve for the weight vectors \mathbf{c}_r and \mathbf{w}_r , using the proposed optimization model,

$$\min_{\substack{\mathbf{c} \in \mathbb{R}^N \\ \mathbf{w} \in \mathbb{R}^k}} f(\mathbf{c}; \mathbf{w}) \text{ subject to } \mathbf{w}^T \mathbf{w} = 1$$

where $f(\mathbf{c}; \mathbf{w}) = \|X\mathbf{c} - Y\mathbf{w}\|_2^2$.

The weight vectors, \mathbf{c}_r , and the survey responses, X_r , are used to develop eight identification (ID) numbers, m_r , for each resource. The ID numbers are developed by finding the average of the product of the survey responses corresponding to students who selected resource r as their preferred resource, and the corresponding weight vector: $m_r = X_r \cdot \mathbf{c}_r$ (Appendix B.6).

The last phase of the validation process to test the model using the survey responses from the test group, $X_{probe} = \begin{bmatrix} -\mathbf{x}^{(1)} - \\ \vdots \\ -\mathbf{x}^{(30)} - \end{bmatrix}$. The predicted resources, \tilde{Y} , are determined by finding the product of X_{probe} and all the derived \mathbf{c} vectors. Then we compute the difference between those results and the ID numbers for each resource, m_r . Lastly, we look at the difference between each product, $X_{probe} \cdot \mathbf{c}_r$, and the corresponding ID number. The resource that contributes to the minimum absolute difference is the predicted resource,

$$\min\{|X_{probe} \cdot \mathbf{c}_1 - m_1|, |X_{probe} \cdot \mathbf{c}_2 - m_2|, |X_{probe} \cdot \mathbf{c}_3 - m_3|, |X_{probe} \cdot \mathbf{c}_4 - m_4|\}.$$

The accuracy of the model will be measured by comparing the predicted resources, \tilde{Y} , and the true ranked resources, Y (Appendix B.7). As illustrated in Figure 5, this process is repeated 9 times, allowing for every group to be tested.

The type of k -Fold Cross Validation method employed here is commonly known as the *Holdout Method*. That is, one data point is left out for testing each time while the remaining data points are used for training the model. This allows us to reduce the variance in the data since the majority of the data is being used to train the model [11].

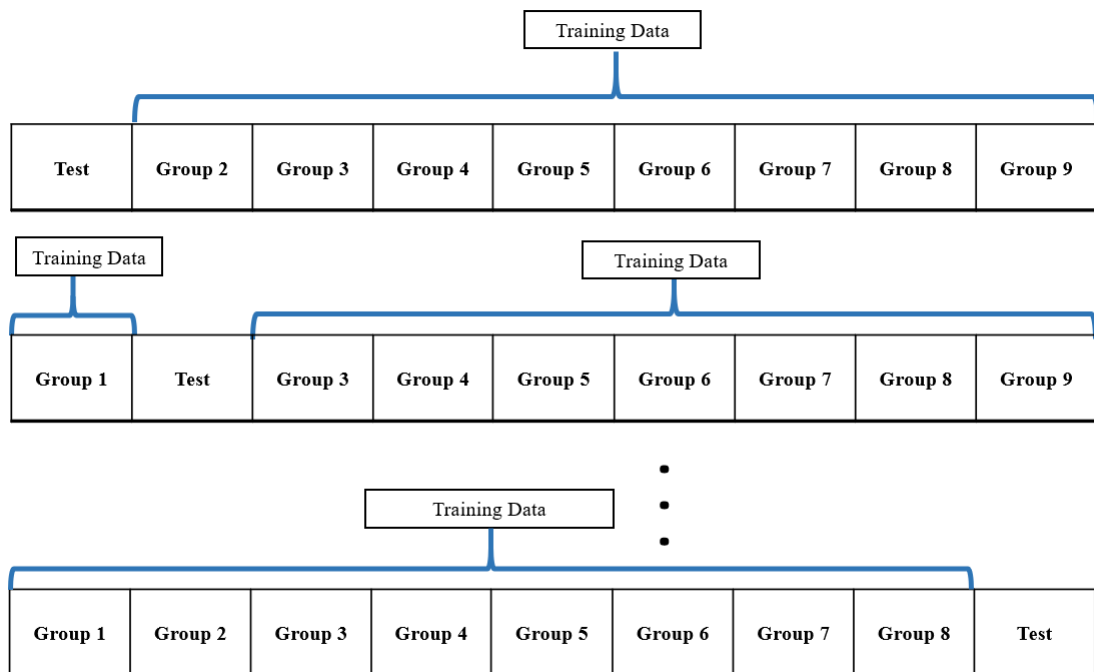


FIGURE 5. A Visual Presentation of the k -Fold Cross Validation method.

CHAPTER 4

EXPERIMENTAL RESULTS

College Student Inventory Survey

The CSI Survey containing 73 questions from the Ruffalo Noel Levitz Form C and 35 questions from Forms A, B, and custom made (Appendix A.1 and A.2), consisted of a total of 108 questions corresponding to 24 categories, shown in Appendix A.3. Notice that question 11 in Appendix A.1 was omitted because it had more than one possible response. An intentional redundancy exists within the survey to increase survey validity. That is, there are multiple questions within each category designed to assess similar information. It is likely that there exists redundancy among different categories as well. For example, *Financial Security* and *Receptivity to Financial Guidance* might capture similar information. *Study Habits* and *Study Skills* might also capture similar information. Having such redundancies can result in lengthy surveys.

As mentioned earlier, surveys of this size are time-consuming and difficult to be administered in a limited-time environment. Thus, our goal is to determine the minimum number of questions needed to obtain similar results as the CSI Survey.

CSULB Data: RNL Form A and C

The TRiO Student Support Service Program (SSSP) administered the CSI Survey containing 108 questions from the RNL Forms A and C to 378 students in the summer of 2016. Three Hundred of those data points were considered usable, meaning we had complete responses for those students.

The students surveyed were incoming freshmen that had similar demographics of students who are considered to be at-risk for not completing their degree. The risk factors include socio-economic status, technology skills, first generation college student, minority group, financial constraints, self-confidence, etc [5]. The students surveyed were offered resources to help their transition from high school to college and help them do well during

their first year.

The questions from the survey assess the students' perceived notions of various at-risk characteristics. We analyzed this data set and used the information gleaned from the analysis to create a reduced survey instrument for the proposed decision aid.

Reduced Survey

Principal Component Analysis was used to determine the number of questions needed in the reduced survey. We empirically chose to retain 44% of the data variance in the reduced survey by studying the singular value distribution shown in Figure 6. Other variances were also analyzed, such as 55% and 50% which were captured with 21 and 18 singular values respectively. When FA was employed using those number of questions, the questions were not grouped appropriately. Thus the determined the numerical rank of the data was 14.

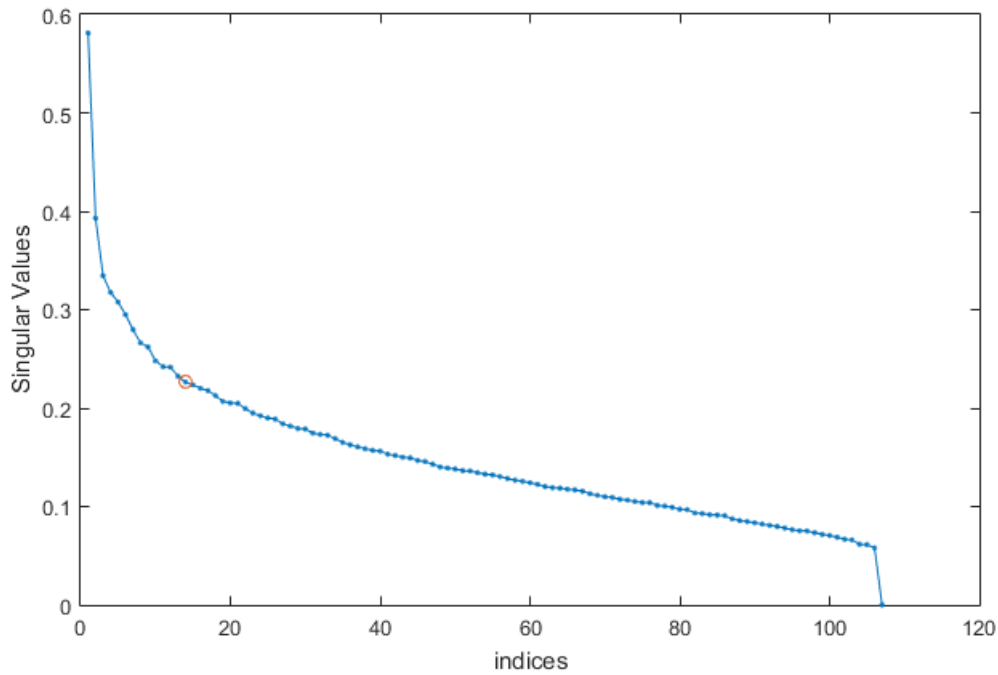


FIGURE 6. The Singular Values of the ESM data gathered.

Factor Analysis was employed to the TRiO SSSP data to determine the dominating

questions for the new survey. The questions from the RNL forms correspond to the 24 categories seen in Appendix A.3. It was determined that using 14 questions to capture 44% of the variances would be best for the reduced survey because the specific variance for the questions was smaller and there was no redundancy.

For example, applying FA to the data set gave us the loadings and the specific variance to the question as follows: I need help improving my study skills.

$$\begin{aligned} x_{61} = & 0.0714f_1 + \mathbf{0.9455}f_2 - 0.0389f_3 + 0.0204f_4 - 0.0053f_5 + 0.0083f_6 - 0.1239f_7 \\ & + 0.1003f_8 - 0.2483f_9 - 0.0334f_{10} - 0.1001f_{11} + 0.0445f_{12} + 0.0472f_{13} \\ & - 0.0133f_{14} + \mathbf{0.3318} \end{aligned}$$

Notice that the highest loading in this question is 0.9455 corresponding to the second factor. This also happens to be the question with smallest variance of 0.3318 corresponding to that factor; therefore this question is representative of the kind of characteristics Factor 2, *Receptivity to Academic Assistance*, assesses. The remaining 13 questions and categories were determined this way and the final reduced survey, *Providing Academic Services for Students* (PASS), can be seen in Appendix A.5.

Decision Aid Data

The decision aid, given in Appendix A.4 and A.5, consists of a list of eight resources with their descriptions and the PASS Survey. It was given to a subset of the incoming freshmen at CSULB during the Early Start Mathematics (ESM) programs in 2017. Students were asked to complete the PASS Survey and, using the descriptions of the eight resources, they were asked to rank them in their preferred order. Preferred order is defined as the order of which they think they would benefit the most from. For example, the first three resources pertain to financial aid assistance, Math tutoring, and Learning Skills Specialists, respectively, thus a ranked response vector would be:

$$\begin{bmatrix} \text{Math Tutoring} \\ \text{Financial Aid} \\ \text{Learning Skills} \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix}$$

where Math tutoring is the most preferred resource.

Students in the ESM programs were working to fulfill the college-readiness requirement. The resources in the decision aid were specifically chosen to help this particular population. From the students surveyed, 162 data points had complete information and they were used in our study.

Classification

The student ranked resources were correlated to the RNL questions using a linear combination of self-efficacy, perceived notions in confidence, and their response to student services. The goal of this study is to recommend the most suitable support service to students based on their responses to the reduced survey. There are two stages for this work: the first stage consists of solving for the weight vectors \mathbf{c} and \mathbf{w} using the training data set and the second stage uses those weight vectors to determine the preferred resource of a participant. The results are compared against the ground truth to determine the accuracy of the model.

Due to instability in solving the eigenvalue problem, only the first four resources were used in the model. In other words, the analyzed responses to the 14 questions of the reduced survey correspond to the participants who selected resources one through four as their most preferred resource. Thus, 126 responses were used in this study.

Implementing the *Holdout Method* implies that the model was trained using 125 survey responses. The student responses were separated by the students' preferred resource, creating four classes, $r = 1, \dots, 4$. The classification pattern for each resource, m_r , was developed by finding the average of the product of the survey responses in each class, X_r , and the derived weight vector, \mathbf{c}_r , for such class. The results from this first stage are four

weight vectors and four classification patterns, each corresponding to one of the resources.

For example, 37 students selected the first resource as their preferred resource. The weight vectors, \mathbf{w}_1 and \mathbf{c}_1 were solved by minimizing the function

$$f(\mathbf{c}; \mathbf{w}) = \left\| \underbrace{\begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,14} \\ x_{2,1} & x_{2,2} & \dots & x_{1,14} \\ \vdots & & \ddots & \vdots \\ x_{37,1} & x_{37,2} & \dots & x_{37,14} \end{bmatrix}}_{X_1} \underbrace{\begin{bmatrix} c_1 \\ \vdots \\ c_{14} \end{bmatrix}}_{\mathbf{c}_1} - \underbrace{\begin{bmatrix} y_{1,1} & \dots & y_{1,4} \\ y_{2,1} & \dots & y_{1,4} \\ \vdots & \ddots & \vdots \\ y_{37,1} & \dots & y_{37,4} \end{bmatrix}}_Y \underbrace{\begin{bmatrix} w_1 \\ \vdots \\ w_4 \end{bmatrix}}_{\mathbf{w}_1} \right\|_2^2.$$

The classification pattern for the first resource, m_1 , is derived from the average of the product of X_1 and \mathbf{c}_1 .

The second phase begins by multiplying the probe responses to the reduced survey, X_{probe} , to each of the weight vectors \mathbf{c}_r , resulting in four scalar products. Then we find the absolute difference between each product and the ID number for each resource. Lastly, the predicted resource, \tilde{Y} , corresponds to the resource, r , whose ID number provides the minimum difference. A visual of the first trial of the second stage is

$$\min\{|X_{probe} \cdot \mathbf{c}_1 - m_1|, |X_{probe} \cdot \mathbf{c}_2 - m_2|, |X_{probe} \cdot \mathbf{c}_3 - m_3|, |X_{probe} \cdot \mathbf{c}_4 - m_4|\}.$$

Suppose the results of the first trial were that $|X_{probe} \cdot \mathbf{c}_2 - m_2|$ was the minimum. The predicted resource for the participant would be $\tilde{Y} = 2$, Math tutoring.

The accuracy of the model was determined by comparing predicted resources to the true ranked resources of the students, $\tilde{Y} - Y$. The smaller the difference between Y and \tilde{Y} , the better the model performs.

Out of the 126 student participants, 32 were accurately matched with their preferred resource. This implies that our model is 29.65% accurate. This is an improvement compared to randomly recommending a resource to students, which would result in 25% accuracy.

CHAPTER 5

DISCUSSION AND FUTURE WORKS

There are several factors that could potentially affect the validity of model.

Availability of Data

Having just 162 survey respondents was not enough to fully train our model because we did not have at least 14 students select each resource as the preferred resource. That is, each resource needs to be trained using responses from a minimum 14 students. A minimum of 120 survey respondents, 15 respondents for each of the resources, are needed to learn the variance of each of the 14 questions. The data acquired did not satisfy these conditions therefore increasing the amount of survey respondents could improve of the likeliness of satisfying the conditions.

The insufficient amount of data points caused issues in solving the eigenvalue problem. The systems used to solve the eigenvalue problem for Resources 5 through 8 were under determined. The responses from students who selected Resources 5 through 8 as their preferred resource were removed from the data, leaving our sample size to 126 students. This study was able to provide recommendations using Resources 1 through 4.

Measure of Fit

The constraint $\|\mathbf{w}\|_2^2 = 1$ was not necessarily a realistic one; instead, $\|\mathbf{w}\|_1 = 1$ would imply that entries of \mathbf{w} represent the percentage of preference for each resource. A benefit of constraining the optimization problem with the L_1 norm is being able to interpret the coefficient values. The L_2 norm forces the coefficients to be similar to each other due to the higher λ values needed to minimize the problem. The L_1 norm would allow for the sparsity in the model and interpret-ability [12].

Quality of Data

Students' perspectives change as they gain experiences and go through different stages of life. The environment and the time at which the students are administered the survey

affects their responses. The survey was administered again in November 2017 to students in Pre-Calculus and Algebra courses. Most of the students were freshmen and only had two months of college experience. The differences in the results between both populations were minimal.

The results of the preferred resources for both populations of students surveyed are seen in Table 3. Notice that the results from the November population are consistent with the results of the summer. This was surprising since a change was expected due to a possible shift in students' priorities after the semester started. The consistency in the responses may show that students do not utilize campus in their first semester.

TABLE 3. Percent of Students That Selected Each Resource

Resources	% of Summer Students	% of November Students
(1) Financial Guidance	23	21
(2) Math Tutoring	32	35
(3) Learning Skills	11	13
(4) English Tutoring	12	13
(5) Career Guidance	7	6
(6) Academic Counseling	7	6
(7) Leadership Skills	6	4
(8) Academic & Social Network	2	2

Another observation is that the first four resources are the most popular. These resources are most commonly known amongst students before they begin at an institution therefore students may be more likely to select them. The decision aid can be improved by providing more information about the resources that are not as common. The uncommon resources could also be shown first in the decision aid to encourage students to read through the descriptions before selecting a resource they are familiar with.

Decision Aid Improvements

The decision aid consisted of eight resources that were considered to be the most appropriate for the population surveyed. The least preferred resource was academic and

social network. Students can be given a different option such as being paired with a peer mentor.

A different population may be studied, such as major specific students, and the model may be trained to determine resource recommendations for the specific population. For example, students in the College of Engineering have resources available only to their students, such as *MAES: Latinos in engineering and Science*. The decision aid could be specialized to address the needs of students in specific colleges to provided more personalized recommendations.

Lastly, a follow-up survey can be incorporated in a future study to follow progress on the use of student resources. Students participating in the training process of the model gain information from the decision aid. They learn about resources their institution has to offer before attending. The descriptors provided help them select resources that would benefit them. A follow-up survey can capture outcomes of using the resources, and it can be used as a tool to gain feedback on the effectiveness of the decision aid.

Survey Analysis

As mentioned above, we were able to train the model to provide recommendations for the first four resources seen in Appendix A.4. Students' responses were separated by their most preferred resource, resulting in four matrices. PCA and SVD were used to determine the amount of questions needed to retain 85% of the statistical variance from the reduced survey, seen in Appendix A.5. FA was employed to attain the significant questions needed to match students to the specific resource. This process can be found in Appendix B.8.

Resource 1: Financial Guidance

Eight questions are needed to retain 85% of the statistical variance. The question with the highest loading for this resource was Question 8: *Please rate your level of agreement to the following statements: My financial obligations are very distracting*. It was expected that this question would have the highest loading because it asks about the students' financial

obligations. The model had 62.3% accuracy in recommending this resource to students.

An interesting observation is that the question with the third highest loading was Question 10: *Please rate your level of agreement to the following statements: My family understands and respects my feelings about most things.* This question tells us that students may be feeling stress from their parents regarding their financial obligations.

Resource 2: Math Drop-in Tutoring

Nine questions are needed to capture 85% statistical variance for Resource 2. It was surprising to see that of those nine questions, the question regarding students' math skills, Question 11, had the third highest loading, not the first. The question with the highest loading was Question 13: *Please rate your level of agreement to the following statements: Most educators are more concerned about themselves than their students.* This tells us that students may feel they need tutoring because they may not feel comfortable communicating with their instructors. We may also infer that students may prefer receiving help through tutoring if they don't feel their instructor cares about them. The model was 40.4% accurate in recommending students this resource.

Resource 3: Learning Skills

Six questions are needed to attain 85% variance for Resource 3. The question with the highest loading was Question 9: *Please rate your level of agreement to the following statements: I would like to learn how to weigh the advantages and disadvantages of various careers.* This was not expected; we expected Question 9 to capture a large amount of variance for Resource 5. Responses to Question 6 were expected to be significant for recommending students this resource. Surprisingly, this question was not considered to be significant.

Only two out of the 17 students, 11.8% that selected Resource 3 as their preferred resource were matched correctly. We can conclude that the survey can be improved by selecting a more appropriate question(s) to capture information regarding Learning Skills.

Resource 4: English Tutoring

Of the seven questions needed to capture 85% variance for this resource, none of them asked the students about their English skills. The model was the least accurate in recommending students Resource 4 to the appropriate students. The model was 10% accurate and mismatched most students to Resource 1.

Concluding Ideas

The model was the most accurate in recommending Resources 1 and 2 to the appropriate students. Of the 48 correctly matched students, 44 of them selected either Resource 1 or 2 to be their preferred resource. Questions may be added or replaced in the survey to capture information that improve the accuracy of the model for Resources 3 and 4. The descriptions of the resources could be improved by providing more details for the unfamiliar resources.

CHAPTER 6

SUMMARY AND CONCLUSIONS

A decision aid was developed in this study to recommend resources for students based on learned individual characteristics. Questions from the known Ruffalo Noel Levitz Forms were analyzed with Principal Component Analysis and Singular Value Decomposition to reduce the number of questions in the survey. Factor Analysis was used to determine the specific questions needed to form a reduced survey, the first instrument of the decision aid.

The reduced survey along with a list of descriptors to eight resources was administered as a decision aid to students in the Early Start Math programs at California State University, Long Beach, during the summer of 2017. The gathered data was used to train and test the proposed model linear constrained optimization model,

$$\min_{\substack{\mathbf{c} \in \mathbb{R}^N \\ \mathbf{w} \in \mathbb{R}^k}} f(\mathbf{c}; \mathbf{w}) \text{ subject to } \mathbf{w}^T \mathbf{w} = 1$$

where $f(\mathbf{c}; \mathbf{w}) = \|X\mathbf{c} - Y\mathbf{w}\|_2^2$, $X_{P \times N}$ was the matrix of the responses to the reduced survey, $Y_{P \times k}$ was the matrix of the participants' ranked resources, and the parameters \mathbf{c} and \mathbf{w} were the corresponding weight vectors. The parameters learned were used to recommend resources to students. The accuracy of the model was determined through the use of the specific k -Fold Cross Validation method, the Holdout Method.

The results show us what types of resources students are interested in and help us develop an improved list with more appropriate resources for future works. The low accuracy of the model tells us that the model can be improved by improving its stability. For example, the large condition numbers, $\|(A - B)^{-1}\| \cdot \|A - B\|$, for each resource, 4, showed that the model was not stable. The decision aid may be improved by adding more questions or obtaining more data points in order to improve the recommendations and stability.

The model can be improved by imposing the L_1 measures on the objective and constraint functions in the optimization problem to gain interpret-ability of the coefficients. It

is also necessary to survey a larger sample size to prevent instability when solving the eigenvalue problem.

TABLE 4. Condition Numbers of Each Resource

Resources	Condition Number
(1) Financial Guidance	1.6949
(2) Math Tutoring	1.4723
(3) Learning Skills	4.7800
(4) English Tutoring	2.4807

This study is the initial step in helping students make informed decisions regarding their academic success. Students can have the ability to learn about themselves and learn about available resources through an improved decision aid developed through future works. The results can be studied by institutions to provide better resources for their students. This can lead to having students make well-informed decisions for themselves as well improving the environment they are in. The ability to succeed academically is in the students' hands.

APPENDICES

APPENDIX A
RNL FORMS AND CSI SURVEYS

A.1 CSI Form C

These are the 73 questions taken from Ruffalo Noel Levitz Form C used by TRiO SSSP at California State University, Long Beach.

Item	Statement	Response options
1.	Class (current term)	1. Freshman; 2. Sophomore; 3. Junior; 4. Senior; 5. Other.
2.	My age category in years is	1. 24 and younger; 2. 25 to 34; 3. 35 to 44; 4. 45 to 54; 5. 55 to 64; 6. 65 and older.
3.	I describe myself as	1. Alaskan Native; 2. American Indian; 3. Asian; 4. Black/African-American; 5. Hispanic or Latino (including Puerto Rican); 6. Native Hawaiian or Pacific Islander; 7. White/Caucasian; 8. Multi-racial; 9. Other
4.	My current marital status	1. Single; 2. Married/domestic partner; 3. Widowed
5.	I support dependents in my household	1. Yes; 2. No
6.	My current enrollment status is	1. Full-time; 2. Part-time
7.	My current level of employment is	1. Full-time; 2. Part-time; 3. Not employed
8.	1. The amount of time I expect to spend working at a job while enrolled in classes	1. 0 (I have no plans to work); 2. 1 to 10 hours per week; 3. 11 to 20 hours per week; 4. 21 to 30 hours per week; 5. 31 to 40 hours per week; 6. More than 40 hours per week
9.	Based on my previous academic performance, I would classify myself as	1. An "A" student; 2. A "B" student; 3. A "C" student; 4. Less than a "C" student
10.	I am the first in my immediate family to go to college	1. Yes (If yes, skip the next item.); 2. No (If no, proceed to the next item and mark all that apply.)
11.	If you responded no to the previous item, select others in your family that have gone to college	1. Spouse; 2. Son; 3. Daughter; 4. Mother; 5. Father; 6. Sister; 7. Brother
12.	My current program of study leads to	1. Associate degree; 2. Bachelor's degree; 3. Master's degree; 4. Doctorate or professional degree; 5. Certification (initial or renewal); 6. Self-improvement/pleasure; 7. Job-related training; 8. Other educational goal
13.	If I could choose, I would complete most of my studies	1. Online; 2. On campus; 3. At a site in my community; 4. At a site outside of my community; 5. At my employment site; 6. Through correspondence courses
14.	My plans at this time are	1. To complete this course/this term; 2. To complete a degree/program at this institution; 3. To take courses to transfer to another institution
15.	I received credit toward my program of study	1. Previous college credits earned; 2. Learning from military training; 3. Learning from prior job or life experiences; 4. More than one above; 5. Other; 6. Not applicable
16.	I made the decision to enroll this term	1. A few days before classes began; 2. A few weeks before classes began; 3. Many months before classes began

Item	Motivational Assessment	Statement
17.	Receptivity to Career Planning	Getting information about the qualifications for various careers would be helpful to me.
18.	Verbal Skills	It is easy for me to figure out the deeper meaning of written material.
19.	Receptivity to Financial Guidance	I need to learn how to manage my finances, including loan and credit card debt.
20.	Study Skills	I generally prefer to study alone.
21.	Receptivity to Academic Assistance	I need help to improve my math skills.
22.	Verbal Skills	I often have difficulty putting my thoughts and ideas into words.
23.	Receptivity to Financial Guidance	I would like to talk with someone about the pros and cons of getting a student loan.
24.	Personal Support	My family does not understand the time I need to spend on my studies.
25.	Life and Career Planning	I am very confused about what occupation is right for me.
26.	Commitment	Taking courses is not the best use of my time right now.
27.	Study Skills	I find it very helpful to participate in study groups.
28.	Personal Support	Family problems often distract me from my studies.
29.	Receptivity to Career Planning	I want to know more about the salaries and opportunities for various careers.
30.	Reading Habits	I only read serious books and articles when I have to.
31.	Use of Technology	I have a weak understanding of how to use computers.
32.	Study Skills	I get so uptight when I study for an exam that I have difficulty concentrating.
33.	Reading Habits	I get a great deal of pleasure from reading.
34.	Study Skills	I am able to balance my schoolwork with obligations at home and work.
35.	Receptivity to Financial Guidance	I would like to talk with a counselor about getting additional financial assistance.
36.	Use of Technology	I find the Internet to be a useful learning tool.
37.	Commitment	I am determined to complete my program of study.
38.	Receptivity to Academic Assistance	Tutoring would benefit me in one or more of my courses.
39.	Verbal Skills	Learning new vocabulary is easy for me.
40.	Study Skills	I have developed a solid system of selfdiscipline that helps me keep up with my studies.
41.	Study Skills	I often feel unprepared for my course assignments.
42.	Receptivity to Academic Assistance	I would like to receive instruction on how to improve my testtaking skills.
43.	Verbal Skills	Speaking in front of others makes me uncomfortable.
44.	Receptivity to Academic Assistance	I want to improve my reading skills.
45.	Life and Career Planning	I have a career action plan that guides my studies.
46.	Attitude Toward Educators	Most educators respect students and treat them fairly.
47.	Life and Career Planning	I would choose the same career, even if my life circumstances were different.
48.	Personal Support	My family understands and respects my feelings about most things.
49.	Receptivity to Career Planning	I need help selecting a career that is right for me.
50.	Reading Habits	Reading has never been one of my favorite pastimes.
51.	Financial Security	Financial problems are not likely to interfere with my studies.
52.	Use of Technology	I seldom rely on the Internet for finding information.

Item	Motivational Assessment	Statement
53.	Receptivity to Career Planning	I would like to learn how to weigh the advantages and disadvantages of various careers.
54.	Commitment	I dread the thought of having to take so many courses.
55.	Math Skills	I have always enjoyed the challenge of trying to solve complex math problems.
56.	Use of Technology	I use a computer to assist me with everyday life and learning.
57.	Commitment	I'm prepared to make the sacrifices needed to reach my educational goals.
58.	Life and Career Planning	I have found an occupation that interests me.
59.	Financial Security	My financial obligations are very distracting.
60.	Life and Career Planning	I fear that my career choice will not pay enough to support the lifestyle I want.
61.	Attitude Toward Educators	Educators tend to have a superior attitude toward students.
62.	Receptivity to Academic Assistance	I need help improving my study skills.
63.	Math Skills	I have difficulty applying even simple math concepts.
64.	Verbal Skills	I can write a clear and wellorganized paper.
65.	Financial Security	I am able to manage my finances without having to work more hours.
66.	Personal Support	My family encourages me to pursue my education.
67.	Math Skills	Math has always been a challenge for me.
68.	Receptivity to Academic Assistance	I need help to improve my computer skills.
69.	Attitude Toward Educators	Most educators are very caring and dedicated.
70.	Commitment	I wonder if my courses are worth all the time, money, and effort I put into them.
71.	Attitude Toward Educators	Most educators are more concerned about themselves than their students.
72.	Receptivity to Academic Assistance	I would like to improve my writing skills.
73.	Commitment	I do not regret the decision to continue my education.

A.2 Custom Codes Survey Questions

These are the 35 added questions from Ruffalo Noel Levitz Forms A & B used by the TRiO SSSP at California State University, Long Beach.

Item	Scale	Source	Statement
1.	[1-4]	Custom	Please rate the degree to which you are comfortable using computer programs as part of your course requirements for general study (eg note taking, organizing, study sheets).
2.	[1-4]	Custom	Please rate the degree to which you are comfortable using computer programs as part of your course requirements for creating documents (eg. Word documents, excel spreadsheets).
3.	[1-4]	Custom	Please rate the degree to which you are comfortable using computer programs as part of your course requirements for creating multimedia presentations (eg/ Powerpoint, Prezi,Vimeo, YouTube).
4.	[1-4]	Custom	Please rate the degree to which you are comfortable using the web to access a 'Course' or 'Learning Management System' (e.g., Beach-Board).

Item	Scale	Source	Statement
5.	[1-4]	Custom	Please rate the degree to which you are comfortable using the web to download or access online audio/video recordings of supplementary content material.
6.	[1-4]	Custom	Please rate the degree to which you are comfortable using the web to access University based services (e.g. enrollment, pay fees).
7.	[1-4]	Custom	Please rate the degree to which you are comfortable using the web to look up or search for information (e.g. online dictionaries, research libraries).
8.	[1-4]	Custom	Please rate the degree to which you are comfortable using the web to send or receive e-mail (e.g. from my instructor, from my classmates).
9.	[1-4]	Custom	Please rate the degree to which you are comfortable using the web for instant messaging to communicate/collaborate with other students in the course.
10.	[1-4]	Custom	Please rate the degree to which you are comfortable using the web to receive alerts about course information (e.g. timetable changes, the release of new learning resources, changes in assessment).
11.	[1-4]	Custom	Please rate the degree to which you are comfortable using the web to upload and share photographs or other digital files related to your course.
12.	[1-4]	Custom	Please rate the degree to which you are comfortable using the web to communicate with the course instructor.
13.	[1-4]	Custom	Please rate the degree to which you are comfortable using the web to manage my time as a personal organizer (e.g. calendar, address book).
14.	[1-4]	Custom	Please rate the degree to which you are comfortable using BeachBoard for viewing my coursework and interacting with my class.
15.	[1-5]	CSI-A	Please rate the degree to which the following statement reflects you, "During the upcoming semester, I expect to feel somewhat lonely and to have a strong desire to see more of my friends and family."
16.	[1-5]	CSI-A	Please rate the degree to which the following statement reflects you, "my life at college is (or will be) quite different from what I'm used to, and the adjustments will be very hard for me to make."
17.	[1-5]	CSI-A	Please rate the degree to which the following statement reflects you, "Over the years, I have frequently been selected as a spokesperson or group leader."
18.	[1-5]	CSI-A	Please rate the degree to which the following statement reflects you, "I often choose or volunteer to be group leader."
19.	[1-5]	CSI-A	Please rate the degree to which the following statement reflects you, "I would like to grow my leadership skills."
20.	[1-5]	CSI-A	Please rate the degree to which you agree with the following statement, "I feel confident of my own opinions, and I'm willing to act on them."
21.	[1-5]	CSI-A	Please rate the degree to which you agree with the following statement, "I like to make my own decisions, and I have a lot of trust in my judgment."
22.	[1-5]	CSI-A	Please rate the degree to which you agree with the following statement, "I often take the initiative in solving my own problems."
23.	[1-5]	CSI-A	Please rate the degree to which you agree with the following statement, "I often get confused when trying to reach major decisions, and I seek a lot of help with them."

Item	Scale	Source	Statement
24.	[1-5]	CSI-A	Please rate the degree to which you agree with the following statement, "On controversial issues, my opinions are often strongly influenced by what other people think."
25.	[1-5]	CSI-B	Please rate the frequency you take very careful notes during class, and review them thoroughly before a test.
26.	[1-5]	CSI-B	Please rate the frequency you study very hard for my courses, even those you don't like.
27.	[1-5]	CSI-B	Please rate the degree to which you agree with the following statement, "I have a very good grasp of the scientific ideas I've studied in school."
28.	[1-3]	CSI-B	Please rate your level of consideration to talk with a counselor about eliminating an unwanted habit (involving food, drugs, cigarettes, or alcohol, etc.)
29.	[1-3]	CSI-B	Please rate your level of consideration to talk with a counselor about some difficulties in my personal/family relationships or social life.
30.	[1-3]	CSI-B	Please rate your level of consideration to talk with someone about getting a scholarship.
31.	[1-3]	CSI-B	Please rate your level of consideration to talk to someone about internship or research position opportunities available.
32.	[1-3]	CSI-B	Please rate your level of to talk to someone about getting a part-time job.
33.	[1-3]	CSI-B	Please rate your level of interest to attend an informal gathering where you can meet some new friends.
34.	[1-3]	CSI-B	Please rate your level of interest to find out more about student government and the various student activities on campus.
35.	[1-3]	CSI-B	Please rate your level of difficulty when organizing your ideas in a written paper, and when avoiding punctuation and grammar mistakes.

A.3 RNL Question Categories

The motivational categories and their corresponding questions in the RNL Form C and custom questions. The bolded questions were used in the reduced survey.

Motivational Assessment	No. of Questions	Item(s)
Personal Demographics	16	CSI-C: 1,2,3,4,5,6, 7 ,8,9,10,11,12,13,14,15,16
Receptivity to Academic Assistance	7	CSI-C: 21, 38, 42, 44, 62 , 68, 72
Commitment	5	CSI-C: 26, 37, 54, 57, 70
Attitude Toward Educators	4	CSI-C: 46, 61, 69, 71
Receptivity to Financial Guidance	3	CSI-C: 19, 23, 35
Financial Security	3	CSI-C: 51, 59 , 65
Life and Career Planning	5	25, 45, 47, 58 , 60
Math Skills	3	CSI-C: 55, 63, 67
Math and Science	1	Custom: 27
Personal Support	4	CSI-C: 24, 28, 48 , 66
Reading Habits	3	CSI-C: 30, 33, 50
Study Skills	6	CSI-C: 20, 27, 32, 34, 40, 41
Study Habits	2	Custom: 25 , 26
Use of Technology	7	CSI-C: 31, 36, 52, 56; Custom: 1, 2, 3
Use of Web	10	Custom: 4, 5, 6, 7, 8, 9, 10, 11, 12, 13
Use of Beachboard	1	Custom: 14
Verbal Skills	5	CSI-C: 18, 22, 39, 43, 64
Verbal Confidence	1	Custom: 36
Receptivity to Career Planning	4	CSI-C: 17, 29, 49, 53
Ease of Transition	2	Custom: 15, 16
Leadership	3	Custom: 17, 18 , 19
Self-reliance	5	Custom: 20, 21, 22, 23, 24
Personal Counseling & Receptivity	5	Custom: 28, 29, 30, 31 , 32
Social Enrichment	3	Custom: 33, 34, 35

A.4 A List of Resources

Survey participants were asked to sort the following resources in the order of which they are likely to utilize. The exact message in the survey reads “The following on-campus resources are available to help you reach your career and academic goals. Please rearrange the resources based on the order for which you would like to use them. Read all of the options before rearranging.”

1. You can speak to a financial aid counselor about the financial aid application process, your eligibility, and your awards. You can also talk to the counselors about the types of loans you can take out, payment plan, and interest rate, etc.
2. You can receive drop-in tutoring in mathematics in one-on-one and group settings at the Learning Assistance Center, where you can get individual help on homework and have the opportunities to work with other students.
3. You can work with Learning Skills Specialists at the Learning Assistance Center to learn challenging material, manage your time more effectively and efficiently, recognize important information in a textbook or lecture, manage stress and anxiety, study for and take tests successfully, and plan research
4. You can work with tutors at the Learning Assistance Center to improve your English speaking, grammar, reading, and writing skills. You can talk with a conversation volunteer to improve your pronunciation, vocabulary, fluency/confidence, and familiarity with American idioms and culture.
5. You can explore different career paths that are suited for your interests and skill sets at the Career Development Center (CDC). Career counselors help you use different factors to make a career decision, such as market demand, physical demand, and job stability. CDC offers free career workshops throughout the year and connects you to potential internships and jobs.
6. You can explore different academic plans and paths with the academic advisors at the University Center for Undergraduate Advising (UCUA). Advisors at UCUA can interpret different academic requirements and policies for you. They can also show you what classes you can take to prepare you for certain careers.
7. You can develop leadership skills by running for officer positions in student clubs. You can also develop leadership skills by attending workshops on public speaking and getting involved with the following centers on campus: Lois J. Swanson Leadership Resource Center, Hauth Center for Communication Skills, and Ukleja Center for Ethical Leadership.
8. You can build a network of faculty and peer support by living and working on campus. The different residential communities allow you to meet students with similar interests as yours and build a network of friends throughout your college career.

A.5 Reduced Survey

The new survey that was created in the research project is given here. It was designed to capture respondents’ perceptions in those motivational categories that exhibited the most variance in the existing TRiO SSSP survey. The questions along with their factor loadings and specific variances are given in the following table. The decision aid messages (Appendix A.4.) were displayed at the end of this survey to gather ground truth information on participants’ preferred ranking of support services they wish to receive. The data collected from this survey was used in training our decision aid model.

Q. No.	Question statement	Factor No.	Factor loading	Specific variance
1.	My current level of employment is	14	-0.8259	0.3603
2.	Please rate the degree to which you are comfortable using computer programs as part of your course requirements for creating multimedia presentations (eg. Pow-erpoint, Prezi, Vimeo, YouTube).	1	0.6408	0.4072
3.	Please rate the degree to which the following statement reflects you: "I often choose or volunteer to be a group leader."	7	0.8215	0.4297
4.	Please rate your level of consideration: To talk to someone about internship or research position opportunities available.	11	0.6166	0.6583
5.	Please rate the frequency you: Take very careful notes during class, and review them thoroughly before a test.	10	0.7208	0.5215
6.	Please rate your level of agreement to the following statements: I need help improving my study skills.	2	0.8934	0.3223
7.	Please rate your level of agreement to the following statements: I have found an occupation that interests me.	3	0.7619	0.5312
8.	Please rate your level of agreement to the following statements: My financial obligations are very distracting.	4	0.7819	0.3814
9.	Please rate your level of agreement to the following statements: I would like to learn how to weigh the advantages and disadvantages of various careers.	5	0.8917	0.3965
10.	Please rate your level of agreement to the following statements: My family understands and respects my feelings about most things.	6	0.7314	0.5111
11.	Please rate your level of agreement to the following statements: Math has always been a challenge for me.	8	0.9370	0.1882
12.	Please rate your level of agreement to the following statements: I can write a clear and well-organized paper.	9	0.6707	0.4521
13.	Please rate your level of agreement to the following statements: Most educators are more concerned about themselves than their students.	12	0.6820	0.4823
14.	Please rate your level of agreement to the following statements: Reading has never been one of my favorite pastimes.	13	0.7629	0.4468

APPENDIX B

MATLAB CODES

B.1 Creating new survey with SVD, PCA, and FA

```
1 clear
2 %% Analysis for CSI-C + Custom Survey (72 + 35 Questions)
3 %% Read files 342 participants X 107 questions
4 % This code determines the number of questions needed for the
5 % reduced survey of the decision aid instrument using PCA, SVD, and FA
6 % We empirically chose to capture 44% of the variance. The principal
7 % components are determined using SVD
8 % Raw_Joined_Data is a file of the survey responses normalized
9 % using the mapping  $\phi = (i-1)/(k-1)$ 
10 Data = xlsread('Raw_Joined_Data',1);
11 %% whitening to normalize data; want data to be in [0,1]
12 [P,Q] = size(Data);
13 Norm_Data = zeros(P,Q); %initialize matrix
14 % whitening removes some of the noise and centers the data at 0
15 for j = 1:Q
16     %finds zscore for EACH column; zscore standardizes data
17     Norm_Data(:,j) = zscore(Data(:,j));
18 end
19 mean_sub = Norm_Data - repmat(mean(Norm_Data,2),[1,Q]);
20 N_D = 1/Q.*mean_sub;
21 %% rescale X
22 [U,S,V] = svd(N_D',0);
23 % finding the variance
24 s = diag(S);
25 k = 0;Ek = 0;
26 Tol = 0.44; % choosing to capture 44% of the variance
27 % variance
28 for i = 1:length(s)
29     var(i) = s(i)^2/sum(s.^2);
30 end
31 % finding rank that satisfies the tolerance
32 while Ek < Tol
33     k = k+1;
34     Ek = Ek + var(k);
35 end
36 Rank = k;
37 %% plotting sigma values
38 figure(1),plot(diag(S),'.-'),%title(['Variance in CSI-C and Custom ...
39     Surveys: Tol ' num2str(Tol),' Rank ' num2str(Rank)])
40 hold on, plot(k,s(k),'o','MarkerSize',6);
41 xlabel('indices'); ylabel('Singular Values')
42 %%%%%%%%% FACTOR ANALYSIS %%%%%%%%%
43 [Loadings, specVar, ~, ~] = factoran(Norm_Data, Rank, 'rotate', 'promax');
44 [Sorted, Indices] = sort(Loadings,'descend');
45 [Sort,Ind] = sort(specVar,'ascend');
46 Qrow18 = cell(1,Rank);
47 Qlabel = cell(Q,Rank);s
```

```

48 %% labeling the questions
49 for col = 1:Rank
50     for r = 1:Q
51         if Indices(r,col) == 0
52             Qlabel{r,col} = [];
53         else if Indices(r,col) ≤ 10
54             Qlabel{r,col} = ['CSI-C ', num2str(Indices(r,col))];
55         else if Indices(r,col) ≥ 11 && Indices(r,col) ≤ 72
56             Qlabel{r,col} = ['CSI-C ', ...
57                             num2str(Indices(r,col)+1, '%3.0f\n')];
58         else
59             Qlabel{r,col} = ['Custom ', ...
60                             num2str(Indices(r,col)-72, '%3.0f\n')];
61         end
62     end
63 end

```

B.2 Script file for classification model

```

1 %% Last Date edited: Jan. 8, 2018
2 % It will separate student responses from ranked resources
3 % It will create a probe and training set of one student
4 % It will solve the Gen Ev problem
5 % It will solve the minimization problem Xc-Yw
6 % It will classify the responses from the test group
7 % It will validate the model
8
9 clear
10 [num,~,~] = xlsread('Pass.Data_sept2');
11 [X_orig,Y_orig] = preparation(num);
12 Full = [X_orig,Y_orig];
13 Full_Sorted = sortrows(Full,15);
14 Valid_Part = Full_Sorted(1:126,:);
15 [g,~] = size(Valid_Part);
16 X_new = Valid_Part(:,1:14); Y_new = Valid_Part(:, 15:end);
17 %% set up for probe and training using keep one out method
18 Resources = zeros(1,g);
19 Resources_true = zeros(1,g);
20 %[Valid_Res] = resource_analysis(Y_orig); % stable resources (More at ...
    least 14 participants)
21 Valid_Res = 4;
22 XC_Box = [];
23 for i = 1:g
24     [XY_sort, X_probe,Y_probe] = ...
25         Probe_Train_Jan1(X_new,Y_new,i,g,Valid_Res);
26     Resources_true(1,i) = Y_new(i,1);
27     %% analyzes data based on each resource j
28     Y_const = 3;% looking at top 3 resources

```



```

28 W = zeros(Y_const-1,Valid_Res);
29 C = zeros(14,Valid_Res);
30 M = zeros(1,Valid_Res); %vector of averages for the 8 resources
31 K = zeros(Valid_Res,g); % number of participants for each resource ...
    8 x X(end)
32 %% separates data points by preferred resource excluding test point
33 cc =1;
34 for j = 1:Valid_Res
35     ind = find(XY_sort(:,15) == j);
36     k = length(ind); % k participants chose resource j
37     K(j,i) = k;
38     X = XY_sort(ind(1):ind(end),1:14); %X matrix for resource j
39     Y = XY_sort(ind(1):ind(end),15:(14+Y_const)); % Y matrix for ...
        resource j
40     Y_norm = zeros(k,Y_const);
41     r = Valid_Res*ones(1,Y_const); % to normalize we have 8 resources
42     for col = 1: Y_const
43         Y_norm(:, col) = (Y(:,col)-ones(k,1))/(r(col)-1);
44     end
45     X_z = zeros(k,14);
46     Y_z = zeros(k,Y_const); %initialize matrix
47     %% whitening removes some of the noise and centers the data at 0
48     for kk = 1:Y_const
49         Y_z(:,kk) = zscore(Y_norm(:,kk)); %finds zscore for EACH ...
            column; zscore finds zero-mean with variance 1
50     end
51     Y = Y_z(:,2:end); % the first col has all j's so the mean is 0
52     for l = 1:14
53         X_z(:,l) = zscore(X(:,l));
54     end
55     rankY(j) = rank(Y);
56     % X_z = X; Y = Y_norm;
57     %% solving the Gen Ev prob to get c and w
58     [c_min,w_min,lambda] = GenEvProb(X_z,Y);
59     %lam(j) = min(lambda)
60     %[c_min,w_min] = minf(c,w,X_z,Y,Y_const);
61     C(:,j) = c_min; % c vectors for the 14 questions
62     W(:,j) = w_min; % w vectors for the valid resources
63
64     [m,XC] = averages(X_z, C,j); %gets average for X_j*c_j
65     XC_Box{j}{i} = XC;
66     M(:,j) = m;
67 end
68 %% matching the resources
69 [Resources] = classification(M,X_probe,Resources,C,i);
70 end
71 Error = Resources - Resources_true;
72 correct = g-nnz(Error);
73 Correct_percent = correct/(g);

```

B.3 Preparing the data for the model

```

1 %% This code cleans and normalizes the data
2 % Inputs: num is the responses of all the students
3 % Outputs: X is the matrix of the student responses to the
4 % survey; Y is the matrix of the student resources
5 function [X,Y] = preparation(num)
6
7 Empty = isnan(num); %finds empty cells NaN; outputs 1's for empty cells
8 [m,n] = size(num);
9 Clean = [] ; %matrix for usable entries
10 S = sum(Empty,2); %if there's a NaN, sum > 0
11 k=1;
12 %% going to fill in Clean matrix using sum of NaN's
13 for row =1:m
14     if S(row) == 0
15         Clean = [Clean;num(row,:)];
16         k = k+1;
17     end
18 end
19 [M,~] = size(Clean);
20 quest = n - 9; %9 resources so n-9 are the questions
21 %% Set up problem
22 X_raw = Clean(:,3:quest);
23 [p,q] = size(X_raw);
24 %% normalizes X with phi mapping
25 X = zeros(p,q);
26 r = [3, 4, 5, 3, 5, 7, 7, 7, 7, 7, 7, 7, 7];
27 % phi = (i - 1) / (k - 1)
28 for col = 1: q
29     X(:, col) = (X_raw(:,col)-ones(p,1))/(r(col)-1); % normalized X
30 end
31
32 Y_old = Clean(:, quest+1:end);
33 [p,~] = size(Y_old);
34 Y = []; % need to delete 9th resource
35 for row = 1:p
36     R = Y_old(row,:);
37     Y(row,:) = R(R<9); % normalized Y
38 end

```

B.4 Separating data into two groups: probe & training

```

1 %% This function separates the data into training and testing participants
2 % Inputs: X_orig & Y_orig are the original normalized survey and
3 % resource responses; i is the participant; g is the total
4 % participants; ValidRes is the amount of valid resources (we can
5 % only solve the ev problem for 4);
6 % Outputs:XY_sort is a training set matrix (144x22) sorted by
7 % preferred resource;X_probe & Y_probe are the probes for the
8 % survey responses groups and resources respectively

```

```

9
10 function [XY_sort, X_probe, Y_probe] = ...
    ProbeTrain_Jan1(X_orig, Y_orig, i, g, Valid_Res)
11     %% probes
12     X_probe = X_orig(i, :); Y_probe = Y_orig(i, :);
13     %% creating training sets
14     F = [1:i-1, i+1:g];
15     p = g - 1;
16     X_train = zeros(p, 14);
17     Y_train = zeros(p, Valid_Res);
18
19     for k = 1:length(F)
20         X_train(k, :) = X_orig(F(k), :); % Responses for all students ...
            minus participant
21         Y_train(k, :) = Y_orig(F(k), 1:Valid_Res); % Resources for all ...
            students minus participant
22     end
23     XY = [X_train, Y_train]; % combines survey responses and matrices to ...
        sort
24     XY_sort = sortrows(XY, 15);

```

B.5 Solving for c & w

```

1  %% This function solves the eigenvalue problem Aw=Bw
2  % Inputs: X is the matrix of responses for the participants that
3  % selected resource j as their preferred response; Y is the matrix
4  % of the students' top three resource responses;
5  % Outputs: c_min & w_min are vectors derived from the
6  % generalized eigenvalue problem. lambda is the lambda
7  % corresponding to the eigenvectors
8
9  function [c_min, w_min, lambda] = GenEvProb(X, Y)
10 A = Y'*Y;
11 B = Y'*X*inv(X'*X)*(X'*Y);
12 %[w, lambda] = eig(A-B);
13 condition_num = norm(A-B, 2)*norm(pinv(A-B), 2);
14 [w, lambda, ~] = svd(A-B);
15 w_min = w(:, end); % the vectors are written corresponding to lambdas in ...
    descending order
16 c_min = inv(X'*X)*(X'*Y)*w_min;

```

B.6 Average of XC; Resource ID Numbers

```

1  %% This function finds the average of matrix XC
2  % Inputs: X_z is the matrix of normalized responses for
3  % participants who selected resource j as their preferred
4  % resource; C is matrix whose column vectors were derived
5  % from the eigenvalue problem and correspond to a specific
6  % j resource

```

```

7 % Outputs: m is the mean of XC and XC is the product of X.z and
8 % the vector
9
10 function [m,XC] = averages(X.z, C,j)
11
12 XC = X.z*C(:,j); %XC product of X for resource j and c.min for resource j
13 m = mean(XC); % averages

```

B.7 Classification of resources for test groups

```

1 %% This function classifies the responses from the test
2 % participant(s)
3 % Inputs: M is a vector of the means  $m^{\wedge}(j)$ , X_probe is the
4 % response of the test participant(s); Resources is the
5 % vector where the predicted ranked resource responses are
6 % stored; C is the matrix of the derived  $c^{\wedge}(j)$  vectors for
7 % each resource; i is the  $i^{\wedge}$ th participant.
8 % Outputs: Resources with the the predicted resource responses for
9 % participants 1 through i.
10 function [Resources] = classification(M,X_probe,Resources,C,i)
11
12 XC = X_probe*C;
13 distance = abs(XC - M);
14 [I,I] = min(distance);
15 Resources(1,i)= I;

```

B.8 Analysis of survey responses for each resource

```

1 % This code performs PCA & SVD on the four matrices
2 % representing the survey responses for the students
3 % selected resources 1 through 4
4 % Inputs: The data collected from the summer ESM students
5 % Outputs: the significant questions needed to match
6 % students to each specific resource
7 clear
8 [num,I,I] = xlsread('Pass.Data.sept2');
9 [X.orig,Y.orig] = preparation(num);
10 Full = [X.orig,Y.orig];
11 Full.Sorted = sortrows(Full,15);
12 Valid.Part = Full.Sorted(1:126,:);
13 [part,I] = size(Valid.Part);
14 X.new = Valid.Part(:,1:14); Y.new = Valid.Part(:, 15:end);
15 g=4;
16 %% set up for probe and training using keep one out method
17 Rank = zeros(1,g);
18 start = 0;Tol = 0.85; % choosing to capture 95% of the variance
19 loads = zeros(15,14); load.ind = zeros(15,14);
20 var = zeros(15,1); var.ind = zeros(15,1);
21 lk=1;

```

```

22 for r = 1:4
23     R = length(find(Y.orig(:,1)==r)); % # of students selected resource i
24     New_pop = ValidPart(1:R,1:14); % response to survey of students
25     Norm_NP = zeros(R,14);
26     for j = 1:14
27         %finds zscore for EACH column; zscore standardizes data
28         Norm_NP(:,j) = zscore(New_pop(:,j));
29     end
30     mean_sub = Norm_NP - repmat(mean(Norm_NP,2),[1,14]);
31     N_D = 1/14.*mean_sub;
32     [U,S,V] = svd(N_D,0);
33     % finding the variance
34     s = diag(S);
35     k = 0;Ek = 0;
36     % variance
37     for i = 1:length(s)
38         var(i) = s(i)^2/sum(s.^2);
39     end
40     % finding rank that satisfies the tolerance
41     while Ek < Tol
42         k = k+1;
43         Ek = Ek + var(k);
44     end
45     Rank(1,r) = k;
46     %%%%%%%%%% FACTOR ANALYSIS %%%%%%%%%%%
47
48     [Loadings, specVar, ~, ~] = factoran(Norm_NP, k, 'rotate', 'promax');
49     [Sorted, Indices] = sort(abs(Loadings),'descend');
50     [Sort,Ind] = sort(specVar,'ascend');
51     loads(lk,1:k) = Sorted(1,:); load_ind(lk,1:k)=Indices(1,:);
52     var(lk,1) = Sort(1,1); var_ind(lk,1)=Ind(1,1);
53     lk = lk+4;
54 end

```

REFERENCES

REFERENCES

- [1] T. Fishman, A. Ludgate, and J. Tutak, *Success by Design: Improving Outcomes in American Higher Education*. Westlake, TX: Deloitte University Press, 2017.
- [2] P. Mills, “High-Impact Practices,” Oct. 2018. [Online]. Available: <https://www.aacu.org/resources/high-impact-practices>.
- [3] V. Montori, “The New Statin Choice Decision Aid,” Nov. 2014. [Online]. Available: <https://shareddecisions.mayoclinic.org/2014/11/19/the-new-statin-choice-decision-aid/>.
- [4] T. Hicks and S. Heastie, “High school to college transition: A profile of the stressors, physical and psychological health issues that affect the first-year on-campus college student,” *Journal of Cultural Diversity*, vol. 15, pp. 143–147, 2017.
- [5] J. Horton, “Identifying at-risk factors that affect college student success,” *Faculty Working Papers from the School of Education*, vol. 7, pp. 83–95, 2008.
- [6] R. Duarte, A. Ramos-Pires, and H. Goncalves, “Identifying at-risk students in higher education,” *Total Quality Management & Business Excellence*, vol. 25, no. 7-8, pp. 944–952, 2014.
- [7] R. N. Levitz, “College student inventory.” [Online]. Available: <https://www.ruffalonl.com/>
- [8] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
- [9] I. Jolliffe, *Principal component analysis*. Springer, 1986.
- [10] H. Mathes and H. Schneeweiss, “Factor analysis and principal components,” *Journal of Multivariate Analysis*, vol. 55, no. 1, pp. 105–124, October 1995.
- [11] T. Fushiki, “Estimation of prediction error by using K-fold cross-validation,” *Statistics and Computing*, vol. 21, no. 2, pp. 137–146, April 2011.
- [12] M. Schmidt, “Least squares optimization with L1-norm regularization,” 2005, CS542B Project Report.