Using Prediction, Clustering, and Diagnostic Models to Improve Teamwork in First-Year Engineering Courses

By
Raphael Jeong-Hin Chin

A thesis submitted in partial fulfillment
of the requirements for the degree of
Bachelor of Science with Honors in Data Science

Department of Statistics
University of Michigan, Ann Arbor
April 18, 2023

ABSTRACT

Team-based learning (TBL) is common in engineering courses, especially in introductory courses and senior design projects. Even though TBL provides students with benefits, not all students are able to benefit equally from the TBL methodology. Coming from vastly different backgrounds (academically and non-academically), students of a team may have different perspectives, aspirations, or personalities. This paper talks about the different models created to predict team performance, clustering students based on their characteristics, and explaining the relationship between the students' latent traits and their characteristics. The dataset that is used in this paper includes student personality inputs, self-and-peer-assessments of teamwork, and perceptions of teamwork outcomes. The data is available via Tandem, a team-support tool developed by the Center for Academic Innovation at the University of Michigan. Using this information, we developed several Bayesian models to predict if a team is working well, several clustering models with the aim of finding ways to group students into teams more effectively, and a few algorithms to predict Q-matrices which are crucial in explaining the relationship between latent traits and students' characteristics in cognitive diagnostic models. These models are able to help faculty members and instructors to gain insights into finding ways to separate students into teams more effectively so that students have a positive team-based learning experience.

Keywords: Bayesian, cognitive diagnostic models, engineering education, Tandem, team-based learning, unsupervised learning, Q-matrix

TABLE OF CONTENTS

ACKNOWLEDGMENTS

Throughout the course of planning and writing this honors thesis, there are many individuals who have graciously dedicated their time and resources. First and foremost, I have to thank my honors thesis advisor, Prof. Gongjun Xu, thesis co-advisor, Dr. Robin Fowler, Jing Ouyang, and Dr. Rebecca Matz. Without their assistance, dedicated involvement, and constructive criticism at every step of the process, this paper would have never been accomplished. I would like to thank them very much for their advice and support throughout the semesters.

I would also like to express my gratitude to Dr. Chenchen Ma, Herbert Li, and Weihan Xu for their support and guidance, which allowed me to set the research direction for this paper. I would also like to thank the Statistics Department and the Tandem team at the Center for Academic Innovation, University of Michigan for providing me with the data and resources needed to complete this paper.

I would like to thank my family and friends for supporting me throughout the semesters when I am preparing for this thesis. Special thanks to Cassandra Goh, Kevin Choong, and Lyana Cremnitz for their willingness to listen to me talk about my research, give me feedback on my ideas, and provide emotional support when I was at a low point. Owing to their support, I was able to stay on track and complete my thesis.

CHAPTER 1. BACKGROUND

In the 1980s, a new learning method, team-based learning (TBL), was first introduced to solve various problems that arose from large class settings [1], [2]. TBL was first implemented in business schools [1], but team-based pedagogy can now be found across engineering, medical, and social sciences programs all around the world. Team-based work can provide students with a wide array of benefits, such as training students to become leaders, providing a platform to give and receive feedback, sharing ideas, and working with others with different backgrounds, both academically and non-academically. In most cases, TBL is useful for students with no working experience since it can serve as a microcosm of the real world and prepare students for the workspace.

Even though TBL provides students with many benefits and opportunities to learn multiple skills, not all students are able to benefit equally from TBL methodology. Since students of a team are usually from different backgrounds, each student may have different perspectives, aspirations, or personalities. For example, in engineering education, it is common for some students in teams to be perceived as not contributing fairly to the team deliverables by the rest of their teammates due to poor performance and low quality of work [3]. Gender has also been found to be related to unbalanced work distribution in engineering teams where women may do more work related to planning or communication, while men may do more technical work [4]. Thus, it is extremely crucial for faculty members and instructors to address these issues properly to create a positive teamwork environment and learning experience for their students.

In order to allow students to reap the benefits of TBL, teamwork assessment and support tools such as Comprehensive Assessment of Team Member Effectiveness (CATME) or Tandem, a team-support tool developed by the Center for Academic Innovation at the University of Michigan [5], can be used to monitor the students' performances and notice any changes within the team [3], [5]–[8]. Some teamwork assessment and support tools are also able to provide real-time insights so that faculty members and instructors can monitor student teams and see the teams' progress. For example, with the help of teamwork assessment and support tools, faculty members and instructors were able to understand how team harmony affects the overall team performance, or how students can be clustered based on their personalities and traits [3], [6]. These tools provide an opportunity for faculty members and instructors to investigate small relationships and trends that might have not been detectable in small samples. Furthermore, teamwork assessment and support tools also allow faculty members and instructors to use student feedback and personalize the help provided.

In this paper, we used data available via Tandem. This paper provides an introduction to using prediction, clustering, and diagnostic models in the field of engineering education. The prediction models provided predictions for students' responses to the "Working Well" item, a variable that indicates how well the team is working together internally (see Fig. 2). For clustering, the models grouped students of similar personalities and traits into the same group based on their responses to the survey questions (see Fig.1). In the diagnostic section, we provided an expert-defined Q-matrix and ways to estimate Q-matrices in engineering education. This paper can be seen as a preliminary guideline in using data collected through the team-support tool to create positive learning experiences not just in the field of engineering education, but also in all fields where teamwork plays an important part of the curriculum.

All of the code used to generate the models used in this paper is available in the paper's git repository [9], and all data used in this paper can be requested through writing submitted to the Center for Academic Innovation. Results from the models were generated through batch jobs within the Great Lakes HPC Cluster provided by ITS Advanced Research Computing at the University of Michigan [10]. Thus, these models are able to handle more data that will be coming in from future semesters, thus making them scalable and reusable.

CHAPTER 2. DATA

This study utilized data collected via Tandem which was first implemented in 2019 and has since collected responses from more than 5000 students. In this paper, data from the "Beginning-of-Term" survey (BoT) and the weekly team check surveys (TC) were studied. These two surveys are described in the subsections under the Data section. The data were studied and used to create the prediction, clustering, and diagnostic models. These models are described in detail in Chapters 3 to 5.

2.1. Surveys

The data used in this paper was from the University of Michigan first-year engineering students. The dataset responses were collected from students enrolled across 14 different sections of an introductory engineering design course, ENGR 100, between Winter 2020 and Fall 2021 (four semesters total). Owing to the COVID-19 pandemic, the latter half of Winter 2020, Fall 2020, and Winter 2021 courses were conducted online or in a hybrid mode. Nonetheless, team-based learning components were present in all courses. We decided to not include the data of the students whose answer to the Gender question was not Male or Female due to the small amount of data present (65 students).

2.1.1. Beginning-of-Term Survey (BoT)

The beginning-of-term survey (BoT) is given to the students at the start of the semester before they have met their course teams. This survey asks about individual characteristics found to be relevant in teamwork literature, such as personality characteristics, previous teamwork experiences, and teamwork preferences. Items in the BoT are based on validated scales in the literature for constructs relevant to teamwork, but to keep the overall survey short, they are single-item and sometimes even double-barreled, based on user testing conducted by the developer of Tandem [5]. 835 BoT survey responses and eight questions from the BoT were used in this study. Students move a slider over seven points for the five questions: "Extraversion", "Procrastination", "Belongingness", "Control", and "SpeakUp". For the remaining three questions ("BT_PastGroups", "BT_PastPositive", and "GroupPreference"), students will choose only one response for each question. These eight questions were chosen as they are most representative of a student's personality and traits. Fig. 1 shows the survey questions and answer choices from the BoT that were used in this study.

2.1.2. Team Check Survey (TC)

The team check survey (TC) is generally given weekly to students and is designed to be mobile-friendly and fast. Students are asked to rate the team (not individuals) overall on five items, which are "working well", "sharing of work", "sharing of ideas", "team confidence", and "logistics/challenges". The dataset consists of 4104 TC survey responses collected from 764 students. Students answered each item on a 9-point Likert scale. In the semesters included in this study, when students responded to one or more of the five items with a 7 or lower (students tend to use only the very top of the scale if they think that their team is doing well), they could additionally select from a list of common teamwork problems the issues that their team was experiencing. All students also were shown an optional text-entry space that they could use to alert instructors regarding issues that their team was facing [5]. Fig. 2 shows the survey questions and answer choices asked in the TC.

Where would you place yourself on the following scales? [7 stops on the scale]			
[Extraversion]	In groups, I tend to listen more than speak.	←→	I often speak up in groups.
[Procrastination]	I usually do work close to a deadline.	←→	I get working on a project as soon as it is assigned.
[Belongingness]	I expect to fit right into the $Course.	←→	I expect to feel pretty out of place in $Course.
[Control]	I think it's good to share work, even if my team might finish tasks differently than me.	←→	I'd rather pick up extra work so I know it's done right.
[SpeakUp]	I'd rather hold back ideas or preferences if my group stays happy.	←→	It's easy for me to speak up about my ideas or preferences even if it disrupts my group.

Where would you place yourself on the following scales? [4 radio buttons]				
	Not at all	Once or Twice	Several Times	Many Times
[BT_PastGroups] Working with a team				

Where would you place yourself on the following scales? [5 radio buttons]					
	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
[BT_PastPositive] My past teamwork experiences were generally positive.					

Where would you place yourself on the following scales? [3 radio buttons]			
	[alone] Work alone	[partner] Work with one partner	[group] Work in a group
[GroupPreference] If given an option, I'd prefer to			

Fig. 1. Snapshot of survey questions and answer choices asked in BoT. "$Course" is replaced by text describing the course (or sometimes, non-course context).

Where would you put your team on each of these scales? [9 stops on the scale]			
[Working Well]	We often have problems working together.	←→	We work really well together.
[Logistics]	We often face logistical barriers (for example, we cannot find convenient meeting times).	←→	We have no problems with logistics (for example, we all stay in touch about the projects).
[Team Confidence]	I worry we don't do well on this project.	←→	We're definitely going to do well on this project.
[Equal Workloads]	The workload is not distributed evenly.	←→	Everyone is pulling their own weight.
[Sharing Ideas]	Only one or two people contribute ideas for our projects.	←→	Everyone evenly contributes ideas for our projects.

Fig. 2. Survey questions and answer choices asked in the TC survey.

2.2. Technology

We used Python and Stan on Google Colab to build all Bayesian models in this paper. For clustering and diagnostic models, we used R and Rstudio. For the diagnostic models, the Jupyter Notebook was also used to run the Restricted Boltzmann Machine-based Q-matrix estimation algorithm.

CHAPTER 3. PREDICTION

3.1. Introduction

In this paper, we created prediction models to predict the students' responses to the "working well" variable, a variable that indicates how well the team is working together internally (see Fig. 2). We define a team to be working well if the "working well" variable is greater than 7 and vice versa. Technically, the models can also predict different variables of interest by changing the inputs provided. For example, instead of "working well", we can also predict "sharing ideas" or "equal workloads" by making other variables in the BoT and TC become independent variables.

3.2. Bayesian Models

The prediction section consists of three different Bayesian models: Simple Logistic Regression, Hierarchical Logistic Regression, and Ordered Logistic Regression. The first two models assume the likelihood function of the Bernoulli distribution due to the binary response variable. The third model was created using the original response variable, "working well", which can hold nine different values. Since the nine different values can be treated as nine different categorical variables, the models assume the likelihood function to be of the ordered logistic distribution. The third model would consider the team to be working well if the predicted score was greater than seven and vice versa. All three models were tested for their accuracy by comparing the predicted response with the actual response collected from the survey. All three models were written using the Stan programming language with help from its documentation [13].

3.2.1. Simple Logistics

The first model assumes that the response variable holds only zeros or ones. Therefore, the likelihood function was designed to be of the Bernoulli distribution as shown in Equation (1). Bernoulli distribution is a discrete probability distribution of a random variable that takes the value 1 with probability p and the value 0 with probability $q = 1 - p$. In this paper, a Bernoulli distributed model with logit parameterization was used because the parameterization would be more numerically stable. The calculation can be simplified [13, Ch. 15.2].

$$y_i \mid x_i \sim Bernoulli(\sigma(\beta^T x_i)), \ \forall i \in \{1, \ldots, n\} \tag{1}$$

where

$\sigma(t) = \frac{1}{1+exp(-t)}$

$\beta \sim Normal(0,\ 2)$

3.2.2. Hierarchical Logistics

A hierarchical logistic regression model was used as the data contain binary response variables and group structures, which in this model refers to the different course sections and genders. Twenty-eight group clusters, formed through the combination of fourteen courses and two genders, were created for this hierarchical logistic regression model. The Cluster Index was calculated using Equation (2) where the male gender has a value of one while the female gender has a value of zero.

$$\text{Cluster Index} = \text{ID}*2 + \text{Gender} - 1 \tag{2}$$

For example, male students in Course 2 will be assigned Cluster Index 4 while female students in Course 2 will be assigned Cluster Index 3. The second model also assumes that the response variable holds only zeros or ones. Therefore, the likelihood function was designed to be of the Bernoulli distribution as shown in Equation (3).

$$y_{ij} \mid x_{ij} \sim Bernoulli(\sigma(\beta_j^T x_{ij}))\ ,\ \forall j \in \{1,\ldots, 28\}\ , \forall i \in \{1,\ldots, n_j\} \tag{3}$$

where

$\sigma(t) = \frac{1}{1+exp(-t)}$

$\beta_j \sim Normal\ (\mu,\ \sigma^2)$

$\mu \sim Normal\ (0, 5)$

$\sigma^2 \sim Uniform(-\ \infty, \infty)$

3.2.3. Ordered Logistics

The third model assumes that the response variable holds values from one to nine. Therefore, the likelihood function was designed to be of the Ordered logistic distribution. The predicted values of this model hold values from one to nine. Then, the predicted values will be converted into ones (if greater than seven, based on the cutoff described in the Team Check Survey section) or zeros (seven or lesser) to be compared with the binary response variable to test the accuracy of the model.

$$\text{OrderedLogistic}(k|\eta,c) = \begin{cases} 1-\text{logit}^{-1}(\eta - c_1) & \text{if } k = 1, \\ \text{logit}^{-1}(\eta - c_{k-1}) - \text{logit}^{-1}(\eta - c_k) & \text{if } 1 < k < K, \text{ and} \\ \text{logit}^{-1}(\eta - c_{K-1}) - 0 & \text{if } k = K \end{cases} \tag{4}$$

where

$\eta = \beta^T x_i$

$\beta \sim Normal(0,\ 2)$

3.3. Results

In order to evaluate and provide statistical inference on the model, the NUTS-HMC sampler was used to produce a set of draws from the posterior distribution of a model conditioned on the training data [14]. HMC-NUTS sampler uses the Hamiltonian Monte Carlo (HMC) algorithm and its adaptive variant, the no-U-turn sampler (NUTS), to produce a set of draws from the posterior distribution of the model parameters conditioned on the data [15]. Each model was trained on 80% of the full data while the remaining 20% was used for testing. The evaluation from the diagnostic statistics is helpful in determining what should be changed in fitting the next models. In the following subsections, some posterior distributions were plotted to check if any of the parameters contain zero within the 94% highest density interval (HDI). The predictive log-likelihood and accuracy will also be used to measure how well each model performs and fit the data. For the ordered logistic model, the predicted values hold values from one to nine. In order to find the accuracy of the model, any value greater than 7 will be treated as one and zero otherwise. The transformed predicted values will then be compared to the true test response.

From Table 1, all three Bayesian models had high accuracy (>75%) in predicting whether the students feel that their teams are working well. These three models also had good performances as the R-hat values were lower than 1.1, meaning that the chains had all converged. Among the 20 different variables, the number of statistically significant variables for Simple Logistics, Hierarchical Logistics, and Ordered Logistics were five, four, and seven respectively. In this paper, we consider a variable to be statistically significant if the posterior distributions of the beta do not contain zero within the 94% highest density interval (HDI) [16].

Among the variables, the four variables that were chosen by all three models were 'TC_Logistics', 'TC_IdeaEquity', 'TC_Workload', and 'TC_Confidence'. These four variables are also the questions asked in the weekly TC surveys distributed to the students. This means that given the BoT and the TC for a particular week, the model is able to inform the instructors if the team is working well or not for that week.

'TC_IdeaEquity_Dir' is the changes (positive, neutral, or negative) in 'TC_IdeaEquity' from the previous to the current week. In the case of the first team check, their values will be zero. Although 'TC_IdeaEquity_Dir' and 'Gender' were not chosen by all three models, it was still chosen by the ordered logistics model suggesting that they are related to the team's work-well score. Additionally, both the logistics and ordered logistics models contain the variable 'Control', suggesting that the variable is related to the team's work-well score as well.

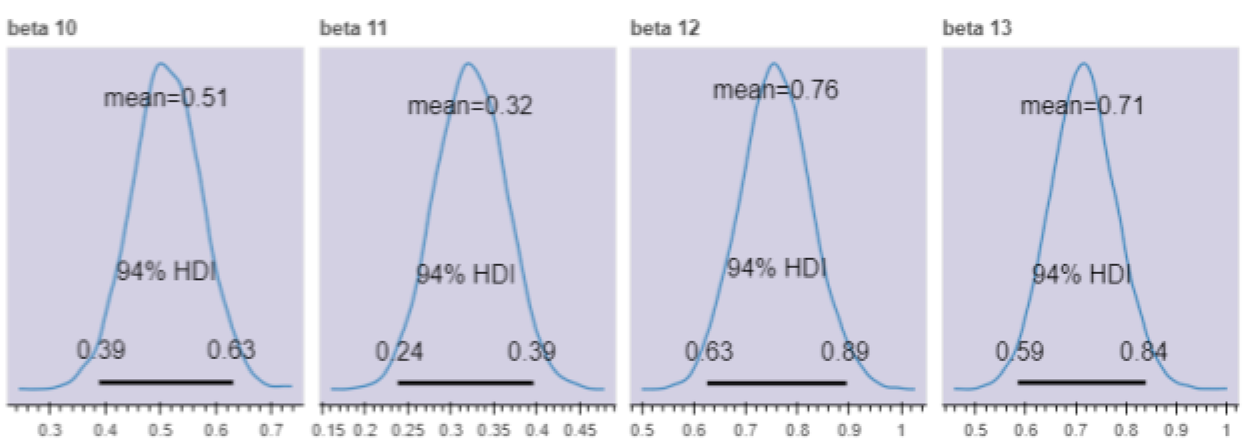

Fig. 3. A portion of the four posterior distributions of the beta ('TC_Workload', 'TC_Logistics', 'TC_Confidence', and 'TC_IdeaEquity') for the Logistics Regression model.

Table 1. Results for the Bayesian models.

Models	Accuracy (%)	R-hat (<1.1)	Divergence	Number of significant variables	Names of variable chosen
Logistics	77.17	TRUE	FALSE	5	['Control', 'TC_Workload', 'TC_Logistics', 'TC_Confidence', 'TC_IdeaEquity']
Hierarchical Logistic	78.89	TRUE	FALSE	4*	['TC_Workload', 'TC_Logistics', 'TC_Confidence', 'TC_IdeaEquity']
Ordered Logistics	77.10	TRUE	FALSE	7	['TC_Workload', 'TC_Logistics', 'TC_Confidence', 'TC_IdeaEquity', 'Control', 'Gender', 'TC_IdeaEquity_Dir']

* Among the 20 different variables, four of them were chosen by more than 60% of the 28 cluster groups.

CHAPTER 4. CLUSTERING

4.1. Introduction

Cluster Analysis is an approach that groups a set of objects such that those objects in the same group are more similar to each other than to those in another group. Although there are a variety of cluster models that can be used, this paper used two different clustering methods that can be easily implemented (hierarchical clustering and K-means clustering) to divide the students into clusters based on how they responded to the eight BoT questions (see Fig. 1). For hierarchical clustering, a dendrogram was plotted and the best number of clusters was chosen based on the plot. For K-means clustering, different methods such as the elbow method and the silhouette method were used to decide the optimal number of clusters.

After clustering, students of the same cluster are more similar to one another compared to students of other clusters. For both clustering methods, the resulting clusters were assigned to the original data, and the statistics for each variable were calculated to show how each cluster has different values for each variable. By clustering the students, faculty members and instructors are able to understand more about the personality distribution of their students. Besides, the clustering information here can be used together with the diagnostic information in the next chapter to understand more about how students in each cluster will respond to the survey questions.

4.2. Hierarchical Clustering

Hierarchical clustering is a clustering method that builds a hierarchy of clusters. The hierarchical clustering used in this paper was agglomerative Hierarchical Clustering. This is a "bottom-up" approach where the process can be summarized as follows:

1. Each observation starts in its own cluster,
2. Pairs of clusters are merged to form a new cluster as they move up the hierarchy.
3. Step 2 is repeated until one cluster is left.

We used the complete-linkage clustering method defined by Equation (5). In this case, we used the Euclidean Distance. This hierarchical clustering process was performed using the *hclust* function from the *stats* package in R [17].

$$max\{d(x,y):\ x \in A,\ y \in B\} \tag{5}$$

where
d(.) is the distance between x and y.

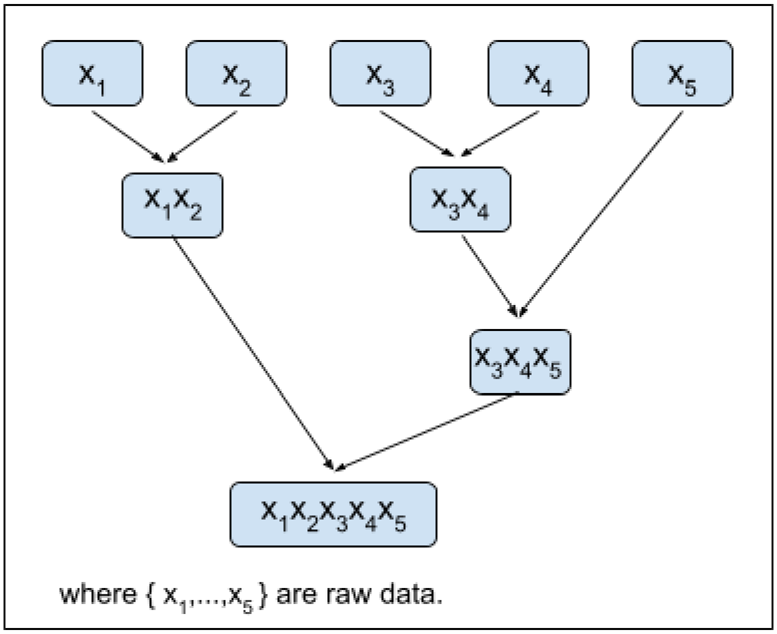

Fig. 4. A simple illustration of the Hierarchical Clustering method.

4.3. K-means Clustering

K-means is a prototype-based, simple partitional clustering algorithm that attempts to find K non-overlapping clusters [18]. The K-means clustering process can be summarized as follows:

1. *K* desired number of clusters is chosen by the user to form the *K* initial centroids.
2. Each data point is assigned to its closest centroid and formed a cluster.
3. The centroid of each cluster will be updated based on new points added to the cluster.
4. Steps 2-3 are repeated until no point changes cluster.

Suppose $D = \{x_1, ..., x_n\}$ is the data to be clustered. Then, the proximities of the data point to the cluster centroids are as follows:

$$\min_{\{m_k\},1\leq k\leq K} \sum_{k=1}^{K} \sum_{x\in C_k} \pi_x \, dist(x, m_k) \tag{6}$$

where

π_x is the weight of **x**,

n_k is the number of data objects assigned to the cluster C_k,

$m_k = \sum_{x \in C_k} \frac{\pi_x x}{n_k}$ is the centroid of the cluster C_k,

K is the number of clusters set by the user.

d(.) is the distance between object **x** and centroid m_k ,$1 \leq k \leq K$.

In this case, we used the Euclidean Distance. This K-means clustering process was performed using the *kmeans* function from the *stats* package in R [17].

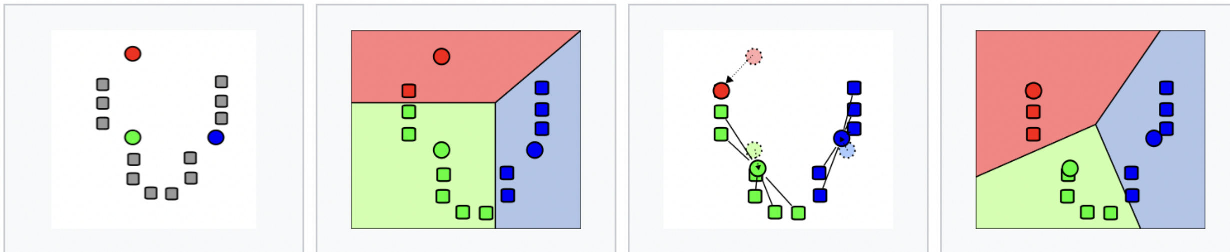

Fig. 5. A simple illustration of the K-Means Clustering method.

4.4. Results

The data was trained with hierarchical clustering and K-means clustering. Based on the result from the clustering algorithms, each data will be assigned a cluster group. For the dataset that was used in this paper, we decided to separate the data into three clusters for both hierarchical clustering and k-means clustering. For hierarchical clustering, the decision to separate the data into three clusters was through looking at the dendrogram that was plotted to ensure that the number of clusters makes sense [19]. For K-means clustering, the decision to separate the data into three clusters was through different algorithms that were designed to help find the number of clusters, such as the elbow method, the silhouette method, and the gap statistic method.

In the elbow method, the cluster value, K, which has the most rapid decline in the sum of squared errors (SSE) is chosen. For the silhouette method, K with a silhouette value closer to 1 is chosen. Silhouette value is the highest when the data points are more compact within the cluster to which it belongs (cohesion) and far away from the other clusters (separation). Lastly, the best K in gap statistics is the one that maximizes the gap statistic value, the change in within-cluster dispersion under the null distribution. The dendrogram plotted from the hierarchical clustering was also used in aiding the selection of the best number of groups [20]–[22].

The total number of assigned groups was six (three from hierarchical clustering and three from K-means clustering). For each group, the mean score for each of the eight BoT variables was calculated and recorded in Table 2. In order to make sure that the mean scores of eight variables recorded in Table 2 were different across clusters of the same clustering method, we used the Kruskal-Wallis rank sum test and pairwise comparisons using the Wilcoxon rank sum test with continuity correction as the post hoc test in our hypothesis test.

From the table, we can see that for hierarchical clustering, students in Cluster 1 tend to have high “Speak Up” scores but low “Belongingness” scores, students in Cluster 2 tend to have high “Procrastination” scores, and students in Cluster 3 tend to have high “Control” scores but low “Extraversion” scores. For K-means clustering, students in Cluster 1 tend to have high “Control” and “Extraversion” scores, students in Cluster 2 tend to have high “Belongingness” scores but low “Speak Up” scores, and students in Cluster 3 tend to have high “BT_PastPositive” scores but low “Control” scores.

Table 2. Results of the clustering models.

	Cluster	Control	Speak Up	Extra version	BT_Past Groups	BT_Past Positive	Group Preference	Procrast ination	Belongingness
Hierarchical Clustering	1	3.64	5.22	4.99	2.69	4.00	2.26	3.61	2.71
	2	3.44	3.68	4.36	2.60	3.81	2.29	4.58	3.47
	3	4.83	3.41	2.72	2.61	3.61	2.12	3.34	3.44
	Cluster	Control	Speak Up	Extra version	BT_Past Groups	BT_Past Positive	Group Preference	Procrast ination	Belongingness
K-means clustering	1	5.04	4.97	5.39	2.74	3.81	2.11	4.41	2.75
	2	3.82	3.24	3.10	2.43	3.69	2.05	3.95	3.95
	3	2.46	4.89	5.10	2.78	4.11	2.59	3.81	2.61

CHAPTER 5. DIAGNOSTICS

5.1. Introduction

The diagnosis section consists of the estimation of Q-matrices and using these Q-matrices to provide insight into the dependency between the variables of BoT and the TC. In this paper, we used the GDINA function from the CDM package [23], [24] to retrieve the delta matrices that are essential to the estimation of the Q-matrices. The initial Q-matrix given to the GDINA function is always $1_{J \times K}$. Both the Lasso and the Truncated L_1 penalty (TLP) terms were used as tuning parameters to retrieve the delta matrices which were then converted to Q-matrices following a similar expectation–maximization (EM) algorithm in [25]. Besides using the GDINA function, we also used the restricted Boltzmann machines (RBMs) algorithm to estimate Q-matrices. For RBMs, the Q-matrices were estimated following a simplified version of the algorithm in [26]. We also used our experience to come up with one expert-defined Q-matrix (Table 4) to compare with the matrices estimated by the models. All the estimated Q-matrices were refined by minimizing the residual sum of squares (RSS) between the real responses and ideal responses using the Qrefine function from the NPCD package [27].

5.2. Cognitive Diagnostic Modeling (CDM)

We used cognitive diagnostic models (CDMs) to understand the relationship between the latent traits that are related to what the TC surveys are characterizing and students' characteristics collected in the BoT. CDMs are psychometric models that provide information about a person's proficiency in solving particular items [28]. We recognize that the survey questions in TCs and BoT do not have correct answers and one does not require any specific proficiency to answer the questions. Nonetheless, CDMs can still be used to capture the relationship between the latent traits that are related to how the students perceive their team experience (questions in the TC) and how the students perceive their own personalities and preferences (questions in the BoT). This motivation can be justified as other studies have used CDMs to learn more about team formation and relationships [29] and between questions in surveys [30].

5.2.1. Q-Matrix

One important component of CDM is the Q-matrix as it contains information on the dependency structure between the J test items and K latent variables [25], [26], Q-matrix can be effectively used to design more intervention strategies. One famous usage of CDMs in the applied world is to study the dependency between mathematical questions and their latent skills for the topic of fractions as shown in Table 3.

Table 3. Q-matrix corresponds to three math questions and three latent attributes.

Questions	Addition	Subtraction	Convert mixed numbers to improper fraction
$2\frac{3}{4} + 1\frac{1}{2}$	1	0	1
$2\frac{3}{4} - 1\frac{1}{2}$	0	1	1
$2\frac{3}{4} - 1\frac{1}{4}$	0	1	0

‘1’ in the Q-matrix means that Skill K is required for the mastery of Item J and vice versa. Thus, Q-restricted latent class models have gained popularity in fields such as educational proficiency assessments, psychiatric diagnosis, and many more disciplines [25]. In this paper, the Q-matrices were either estimated from the GDINA model or defined by experts. Workload, Confidence, and Sharing Ideas are three latent traits of students related to what TC is characterizing. Table 4 shows the Q-matrix defined by the experts.

Table 4. Experts defined Q-matrix, Q0.

Items	Workload	Confidence	Sharing Idea
Control	1	0	0
SpeakUp	0	0	1
Extraversion	0	0	1
BT_PastGroups	0	1	0
BT_PastPositive	0	1	0
GroupPreference	1	1	0
Procrastination	1	0	0
Belongingness	1	1	1

5.2.2. Generalized-Deterministic Inputs, Noisy "and" gate (GDINA) model

The GDINA model assumes a conjunctive relationship among attributes, where it is necessary to possess all the attributes indicated by the Q-matrix to be capable of providing a positive response [25]. The GDINA model requires a $J \times K$ Q-matrix and for each cell of the Q-matrix, q_{jk} is 1 if the k^th^ attribute is required to answer the j^th^ item positively. Nonetheless, GDINA separates the latent classes into $2^{K_j^*}$ latent groups, which $K_j^* = \sum_{k=1}^{K} q_{jk}$ represent the number of required attributes for item j [31]. According to [31], we can let α_{lj}^* be the reduced attribute vector whose elements are the required attributes for item j. For example, if only the first two attributes are required for item j, then the attribute vector α_{lj} reduces to $\alpha_{lj}^* = (\alpha_{lj1}, \alpha_{lj2})'$. Using α_{lj}^* reduces the number of latent groups to be considered for item j from 2^K to 2^κ where $\kappa = K_j^*$. Then the probability that examinees with attribute pattern α_{lj}^* will answer item j correctly is denoted by

$$P(X_j = 1|\alpha_{lj}^*) = P(\alpha_{lj}^*) \qquad (7)$$

Although there are multiple link functions discussed in [31], this paper uses only the identity link function which is given in Equation (8).

$$P(\alpha_{lj}^{*}) = \beta_{j0} + \sum_{k=1}^{K_j^*} \beta_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \beta_{jkk'}\alpha_{lk}\alpha_{lk'} ... + \beta_{j12...K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk} \quad (8)$$

where

β_{j0} is the intercept for item j;

β_{jk} is the main effect due to α_k;

$\beta_{jkk'}$ is the interaction effect due to α_k and $\alpha_{k'}$;

$\beta_{j12...K_j^*}$ is the interaction effect due to $\alpha_1, ..., \alpha_{K_j^*}$

5.2.3. Delta-Matrix

The $J \times 2^K$ delta matrix returned by the function will be converted into a $J \times (2^K - 1)$ binary matrix with the intercept column removed. The idea behind this is that since $\delta = \beta \times q$, if δ is not 0, q is definitely not 0, where β and q are elements in Equation (8). Values that are close to 0 in the delta matrix (smaller than 0.1) will be forced to be 0 and everything else to be 1 as shown in Equations (9) and (10). The $J \times 2^K$ binary matrix will be collapsed into a $J \times K$ binary matrix by grouping up the latent attributes that are required to master the item J.

Let $\alpha \in \{0, 1\}$, $1 \le k \le K$, and $\delta_{ji} = \alpha_{iK} ... \alpha_{i1}$ be the binary representation index of ith element in the jth row of the delta matrix. δ_{ji} will be transformed to have a value of 1 if it is greater than the threshold and 0 otherwise.

$$t_{jk} = \sum_{k=1}^{K} \delta_{ji} \text{ where } \alpha_{ik} = 1 \quad (9)$$

$$\widehat{Q}_{jk} = 1 \text{ iff } t_{jk} \neq 0 \quad (10)$$

For example, let $\delta = (1.4, 1.32, 0.08, 2.1, 0.0003, 0.0001, 0)$, $J = 1$, $K = 3$, and threshold = 0.1, then applying Equation (9), we get,

$$\delta = (1.4, 1.32, 0.08, 2.1, 0.0003, 0.0001, 0) \Rightarrow (1, 1, 0, 1, 0, 0, 0)$$

$$t = (2, 2, 0)$$

Applying Equation (10), we get, $\widehat{Q} = (1, 1, 0)$. In (9), the columns of the $J \times (2^K - 1)$ binary matrix refers to (Attr1, Attr2, Attr3, Attr12, Attr13, Attr23, Attr123). The matrix is then collapsed into a $J \times K$ matrix by summing up all the 1s into their respective latent attributes, where the columns refer to (Attr1, Attr2, Attr3). If $t_{jk} \neq 0$, then it will become 1 as shown in (10).

The estimated Q-matrix in (10) is expected to be identifiable only up to rearranging the orders of the columns. This is because when estimating the Q-matrix, the columns do not contain information about the latent attributes. (e.g. the nth column of the Q-matrix might not refer to the nth latent attribute). Thus, the estimated Q-matrix will be reordered so that each column shows the lowest possible average Tucker index congruent coefficient with the True Q-matrix's columns. This process was done using the orderQ function in cdmTools [32].

Algorithm 1: Q-matrix estimation

Input: $\delta_{J\times J}$, λ

Output: Estimates $\widehat{Q_{J\times K}}$

Initialize t = 0.1

for seed = 1,...,50 **do**

for each penalty term in λ **do**

1. Record the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) of the model.
2. Retrieve the $\delta_{J\times 2^K}$ matrix

end

end

Obtain the models with the lowest five mean AIC and five mean BIC for LASSO and TLP

for each selected model **do**

1. Perform Equations (9) and (10).

end

5.3. Restricted Boltzmann Machines (RBMs)

RBMs are generative stochastic artificial neural network models that can learn probability distributions over a collection of inputs. RBMs were initially invented by Paul Smolensky under the name Harmonium [33]. RBMs used in this paper will follow the model design in [26]. Visible units are denoted by $R = \{R_1,... R_j\} \in \{0, 1\}^J$ and hidden units by $\alpha = \{\alpha_1,... \alpha_j\} \in \{0, 1\}^K$. RBMs are characterized by the energy functions with the joint probability distribution given by:

$$P(R, \alpha; \theta) = \frac{1}{Z(\theta)} exp\{-\ E(R, \alpha; \theta)\} \tag{11}$$

where $Z(\theta)$ is the partition function given by

$$Z(\theta) = \sum_{R\in\{0,1\}^J} \sum_{\alpha\in\{0,1\}^K} exp\{-\ E(R, \alpha; \theta)\} \tag{12}$$

and $E(R, \alpha; \theta)$ is the energy function given by

$$E(R, \alpha; \theta) = -\ b^T R\ -\ c^T \alpha\ -\ R^T W \alpha = -\sum_{j=1}^{J} R_j b_j - \sum_{k=1}^{K} \alpha_k c_k - \sum_{j=1}^{J}\sum_{k=1}^{K} R_j w_{j,k} \alpha_k \tag{13}$$

In Equations 11-13, $\theta = \{b, c, W\}$ are the model parameters, $b \in R^J$ are visible biases, $c \in R^K$ are hidden biases and $W \in R^{J\times K}$ is the weight matrix describing the interactions between the visible and the hidden units. The hidden and visible units are conditionally independent as there are no "R-R" or "α-α" interactions [26].

$$W = \begin{bmatrix} w_{11} & 0 & w_{13} & 0 \\ 0 & w_{22} & 0 & w_{24} \\ w_{31} & 0 & w_{33} & 0 \\ w_{41} & 0 & 0 & w_{44} \\ 0 & w_{52} & w_{53} & 0 \end{bmatrix} \implies Q = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

Fig. 6. Relationship between the weight matrix and the Q-matrix.

On the left-hand side of Fig. 6 is the weight matrix, W, for RBMs. For $w_{j,k} \neq 0$ indicates the presence of interaction between the visible and hidden units. Although the GDINA model violates the conditionally independent assumptions of RBM, it was shown in [26] that the Q-matrices for these models are estimable.

5.4. Results

From the prediction section in Chapter 3, it is observed that the four variables of the TC are the main variables that can be used to predict the outcome of the team's work-well score. Nonetheless, instructors are also interested in understanding why the students choose to answer the four TC questions with high or low scores. We hypothesize the students' TC responses might be related to the student's personal traits and characteristics which can be obtained from the eight BoT questions used in this paper. Table 5 contains the results for the information criterion for the twenty-six estimated Q-matrices.

20 different Q-matrices were estimated using Algorithm 1 and 5 different Q-matrices were estimated using the RBM algorithm. In order to determine the Q-matrix that can best express the relationship between the *J* items and *K* latent skills, the 25 estimated Q-matrices and one expert-defined Q-matrix were accessed again using the GDINA function, and the Q-matrix with the lowest AIC and the lowest BIC were returned. From Table 5, Model 13 has the lowest AIC, and Model Q0 has the lowest BIC.

From Table 5, we can observe that the two preferred Q-matrices are not the same, but they have a lot of similarities. For example, for the variables SpeakUp and Extraversion, both matrices agree that they are related to Sharing Idea, and the variable Procrastination is related to Workload. Lastly, it is important to ensure that both Q-matrices are identifiable because an identifiable matrix is crucial for the consistent estimation of the model parameters of interest and valid statistical inferences [25], [34]. Both the Q-matrices 13 and Q0 are generically identifiable after checking with the identifiability conditions in Theorem 4 of [35, Sec. 5].

Table 5. Results for twenty-one Q-matrices.

Q_matrix	AIC	BIC	Q_matrix	AIC	BIC
1	7435.3	7714.3	14	7432.9	7655.1
2	7437.7	7650.4	15	7442.8	7646.1
3	7432.4	7654.6	16	7432.9	7655.1
4	7437.7	7650.4	17	7431.8	7672.9
5	7432.4	7654.6	18	7432.9	7655.1
6	7432.4	7597.9	19	7430.7	7671.8
7	7431.4	7615.7	20	7447.9	7717.4
8	7431.4	7615.7	21	7440.2	7624.6
9	7431.4	7615.7	22	7449.0	7623.9
10	7427.6	7687.6	23	7486.7	7633.2
11	7450.2	7729.1	24	7476.9	7632.9
12	7449.7	7653.0	25	7539.1	7704.6
13	7426.9	7686.9	Q0	7429.6	7576.2

Table 6. Q-matrix 13 (left) and Q-matrix Q0 (right)

	Attr1	Attr2	Attr3			Attr1	Attr2	Attr3
Control	1	1	1		Control	1	0	0
SpeakUp	1	1	1		SpeakUp	0	0	1
Extraversion	1	1	1		Extraversion	0	0	1
BT_PastGroups	0	0	1		BT_PastGroups	0	1	0
BT_PastPositive	0	0	1		BT_PastPositive	0	1	0
GroupPreference	1	0	1		GroupPreference	1	1	0
Procrastination	1	1	1		Procrastination	1	0	0
Belongingness	1	1	1		Belongingness	1	1	1

Attr1,2,3 are Workload, Confidence, Sharing Idea respectively.

CHAPTER 6. DISCUSSION AND FUTURE DIRECTIONS

For the prediction section, the Hierarchical Logistic Regression model had the highest accuracy. Even though the accuracy is higher, the computational time for that model to run 4 chains is longer (1.5 hours) compared to the Logistic Regression model (5 minutes). The runtime for the Ordered Logistic model (1.75 hours) is similar to those of Hierarchical Logistic Regression. Therefore, the simpler Logistic Regression model is preferred compared to the other models. Some improvements could be made to the models in the future. Penalty terms such as Lasso, Ridge, and Elastic nets can be used to increase the accuracy of the Bayesian models. Since the ordered logistic regression also performed extremely well in estimating the teams' work-well scores, we believe that the bounded discrete distributions might also be another way to predict team outcomes in the future.

For the clustering section, context experts (in this case, engineering faculty and/or engineering students) should make sense of the clusters generated by the models by characterizing them according to the students' characteristics. The data used in this clustering only contained the students' responses to the eight BoT questions and excluded information such as gender, race, and course sections. Thus, the goal would be to have a database of different clusters of students and their associated characteristics as described by their team members so that when new clusters appear as similar to clusters in the database, the instructors would have a more comprehensive understanding of the personalities of students in such clusters to put students into groups that best fit the students.

For the diagnostic section, we were able to obtain two preferred Q-matrices that can best express the relationship between the items asked in BoT and latent skills observed in the weekly TC. In the future, researchers can try to improve on the Q-matrix estimation by using other CDM models such as DINA, DINO, SDINA, or by using other estimation algorithms such as EM stepwise estimation with a provisional Q-matrix [25]. Moreover, context experts should also look at both the estimated Q-matrix and the expert-defined Q-matrix to make sure they make sense.

As students respond to a variety of questions that reveal information about themselves in various surveys, it is a challenge for faculty members and instructors to keep track of each characteristic and observe changes or trends among students in a team. With the help of team formation tools such as Tandem or CATME, faculty members and instructors are now able to use the data collected to understand the team dynamics and provide help if problems arise. This paper is a first step towards using students' responses collected from teamwork assessment and support tools to predict team performance, clustering students based on their characteristics, and explaining the relationship between the students' latent traits and their characteristics. Hopefully, more research in this area will be conducted to ensure students can enjoy positive teamwork experiences.

WORKS CITED

[1] L. K. Michaelsen, W. Watson, J. P. Cragin, and L. Dee Fink, "Team Learning: a Potential Solution To the Problems of Large Classes," *Exch. Organ. Behav. Teach. J.*, vol. 7, no. 1, pp. 13–22, Jan. 1982, doi: 10.1177/105256298200700103.

[2] V. Najdanovic-Visak, "Team-based learning for first year engineering students," *Educ. Chem. Eng.*, vol. 18, pp. 26–34, Jan. 2017, doi: 10.1016/j.ece.2016.09.001.

[3] J. H. Chin, Y. Gao, H. Li, M. P. Su, and R. Fowler, "Predicting Team Project Score: It's More about Team Harmony and Less about Individual Performance," presented at the 2020 ASEE Virtual Annual Conference Content Access, Jun. 2020. Accessed: Oct. 28, 2022. [Online]. Available: https://peer.asee.org/predicting-team-project-score-it-s-more-about-team-harmony-and-less-about-individual-performance

[4] R. R. Fowler and M. P. Su, "Gendered Risks of Team-Based Learning: A Model of Inequitable Task Allocation in Project-Based Learning," *IEEE Trans. Educ.*, vol. 61, no. 4, pp. 312–318, Nov. 2018, doi: 10.1109/TE.2018.2816010.

[5] "Tandem." https://tandem.ai.umich.edu/welcome (accessed Mar. 07, 2021).

[6] J. H. Chin, H. Li, and R. Fowler, "Proof of Concept: An Algorithm for Consideration of Students' Personalities in Team Formation," presented at the 2021 ASEE Virtual Annual Conference Content Access, Jul. 2021. Accessed: Oct. 28, 2022. [Online]. Available: https://peer.asee.org/proof-of-concept-an-algorithm-for-consideration-of-students-personalities-in-team-formation

[7] M. L. Loughry, M. W. Ohland, and D. J. Woehr, "Assessing Teamwork Skills for Assurance of Learning Using CATME Team Tools," *J. Mark. Educ.*, vol. 36, no. 1, pp. 5–19, Apr. 2014, doi: 10.1177/0273475313499023.

[8] R. Fowler, L. K. Alford, S. Sheffield, C. Hayward, T. S. Henderson, and R. L. Matz, "Supporting Equitable Team Experiences Using Tandem, an Online Assessment and Learning Tool," presented at the 2021 ASEE Virtual Annual Conference Content Access, Jul. 2021. Accessed: Feb. 09, 2023. [Online]. Available: https://peer.asee.org/supporting-equitable-team-experiences-using-tandem-an-online-assessment-and-learning-tool

[9] J. H. Chin, "Utilizing Data-Driven Techniques to Improve First-Year Engineering Courses: Using Prediction, Clustering, and Diagnostic Models in Engineering Education." Mar. 27, 2023. Accessed: Mar. 27, 2023. [Online]. Available: https://github.com/jeonghin/HonorsThesisDS2023

[10] "Great Lakes," *ITS Advanced Research Computing*. https://arc.umich.edu/greatlakes/ (accessed Mar. 27, 2023).

[11] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*. Philadelphia, PA, UNITED STATES: CRC Press LLC, 2013. Accessed: Oct. 28, 2022. [Online]. Available: http://ebookcentral.proquest.com/lib/umichigan/detail.action?docID=1438153

[12] "SAS Help Center: Bayesian Analysis: Advantages and Disadvantages." https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.4/statug/statug_introbayes_sect015.htm (accessed Oct. 28, 2022).

[13] S. D. Team, *Stan Functions Reference*. Accessed: Feb. 03, 2023. [Online]. Available: https://mc-stan.org/docs/functions-reference/index.html

[14] "Stan Documentation," *stan-dev.github.io*. //mc-stan.org/users/documentation/ (accessed Dec. 11, 2022).

[15] "MCMC Sampling — CmdStanPy 0.9.64 documentation." https://cmdstanpy.readthedocs.io/en/stable-0.9.65/sample.html (accessed Feb. 07, 2023).

[16] "arviz.summary — ArviZ 0.11.2 documentation." https://oriolabril.github.io/arviz/api/generated/arviz.summary.html (accessed Feb. 04, 2023).

[17] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2021. [Online]. Available: https://www.R-project.org/

[18] J. Wu, “Cluster Analysis and K-means Clustering: An Introduction,” in *Advances in K-means Clustering: A Data Mining Thinking*, J. Wu, Ed., in Springer Theses. Berlin, Heidelberg: Springer, 2012, pp. 1–16. doi: 10.1007/978-3-642-29807-3_1.

[19] F. Nielsen, “Hierarchical Clustering,” in *Introduction to HPC with MPI for Data Science*, F. Nielsen, Ed., in Undergraduate Topics in Computer Science. Cham: Springer International Publishing, 2016, pp. 195–211. doi: 10.1007/978-3-319-21903-5_8.

[20] B. Mirkin, “Choosing the number of clusters,” *WIREs Data Min. Knowl. Discov.*, vol. 1, no. 3, pp. 252–260, 2011, doi: 10.1002/widm.15.

[21] C. Yuan and H. Yang, “Research on K-Value Selection Method of K-Means Clustering Algorithm,” *J*, vol. 2, no. 2, Art. no. 2, Jun. 2019, doi: 10.3390/j2020016.

[22] R. Tibshirani, G. Walther, and T. Hastie, “Estimating the number of clusters in a data set via the gap statistic,” *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 63, no. 2, pp. 411–423, 2001, doi: 10.1111/1467-9868.00293.

[23] A. Robitzsch, T. Kiefer, A. C. George, and A. Uenlue, “CDM: Cognitive Diagnosis Modeling.” Apr. 11, 2022. Accessed: Apr. 12, 2022. [Online]. Available: https://CRAN.R-project.org/package=CDM

[24] A. C. George, A. Robitzsch, T. Kiefer, J. Groß, and A. Ünlü, “The *R* Package **CDM** for Cognitive Diagnosis Models,” *J. Stat. Softw.*, vol. 74, no. 2, 2016, doi: 10.18637/jss.v074.i02.

[25] G. Xu and Z. Shang, “Identifying Latent Structures in Restricted Latent Class Models,” *J. Am. Stat. Assoc.*, vol. 113, no. 523, pp. 1284–1295, Jul. 2018, doi: 10.1080/01621459.2017.1340889.

[26] C. Li, C. Ma, and G. Xu, “Learning Large Q-Matrix by Restricted Boltzmann Machines,” *Psychometrika*, Jan. 2022, doi: 10.1007/s11336-021-09828-4.

[27] Y. Zheng and C.-Y. Chiu, “NPCD: Nonparametric Methods for Cognitive Diagnosis.” 2019. [Online]. Available: https://CRAN.R-project.org/package=NPCD

[28] M. von Davier and Y.-S. Lee, Eds., *Handbook of Diagnostic Classification Models: Models and Model Extensions, Applications, Software Packages*. in Methodology of Educational Measurement and Assessment. Cham: Springer International Publishing, 2019. doi: 10.1007/978-3-030-05584-4.

[29] Y. Liu *et al.*, “Collaborative Learning Team Formation: A Cognitive Modeling Perspective,” in *Database Systems for Advanced Applications*, S. B. Navathe, W. Wu, S. Shekhar, X. Du, S. X. Wang, and H. Xiong, Eds., in Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016, pp. 383–400. doi: 10.1007/978-3-319-32049-6_24.

[30] J. L. Templin and R. A. Henson, “Measurement of psychological disorders using cognitive diagnosis models,” *Psychol. Methods*, vol. 11, pp. 287–305, 2006, doi: 10.1037/1082-989X.11.3.287.

[31] J. de la Torre, “The Generalized DINA Model Framework,” *Psychometrika*, vol. 76, no. 2, pp. 179–199, Apr. 2011, doi: 10.1007/s11336-011-9207-7.

[32] P. Nájera, M. A. Sorrel, and F. J. Abad, “cdmTools: Useful Tools for Cognitive Diagnosis Modeling.” Mar. 30, 2022. Accessed: Apr. 12, 2022. [Online]. Available: https://CRAN.R-project.org/package=cdmTools

[33] P. Smolensky, “Information Processing in Dynamical Systems: Foundations of Harmony Theory,” COLORADO UNIV AT BOULDER DEPT OF COMPUTER SCIENCE, Feb. 1986. Accessed: Apr. 24, 2022. [Online]. Available: https://apps.dtic.mil/sti/citations/ADA620727

[34] Y. Gu and G. Xu, “The Sufficient and Necessary Condition for the Identifiability and Estimability of the DINA Model,” *Psychometrika*, vol. 84, no. 2, pp. 468–483, Jun. 2019, doi: 10.1007/s11336-018-9619-8.

[35] Y. Gu and G. Xu, “Sufficient and Necessary Conditions for the Identifiability of the Q-matrix,” *Stat. Sin.*, 2021, doi: 10.5705/ss.202018.0410.