Predicting Team Function Using Bayesian and Cognitive Diagnostic Modeling Approaches

Mr. Jeong Hin Chin, University of Michigan

Jeong Hin Chin graduated from the University of Michigan College of Literature, Science, and the Arts with a triple degree in Honors Data Science, Honors Asian Studies, and Statistics. He will be joining the University of Michigan School of Information as a Master's student starting Fall 2023. He is interested in clustering methods, cognitive diagnostic models, educational tools, mHealth, and machine learning.

Ms. Jing Ouyang, University of Michigan

Jing Ouyang is a Ph.D. Candidate in the Department of Statistics at University of Michigan, supervised by Prof. Gongjun Xu. Before coming to Michigan, I received a BSc. in Mathematics and Economics from the Hong Kong University of Science and Technology in 2019. Her research interests primarily lie in latent variable models, psychometrics, high-dimensional statistical inference and statistical machine learning. Specifically, she is working on developing statistical theory and methodology to analyze high-dimensional and complex data with latent variables for interdisciplinary research.

Dr. Robin Fowler, University of Michigan

Robin Fowler is a Technical Communication lecturer and a Engineering Education researcher at the University of Michigan. Her teaching is primarily in team-based engineering courses, and her research focuses on equity in communication and collaboration as well as in group design decision making (judgment) under uncertainty. She is especially interested in how power relationships and rhetorical strategies affect group judgment in engineering design; one goal of this work is to to understand factors that inhibit full participation of students who identify with historically marginalized groups and investigate evidence-based strategies for mitigating these inequities. In addition, she is interested in technology and how specific affordances can change the ways we collaborate, learn, read, and write. Teaching engineering communication allows her to apply this work as she coaches students through collaboration, design thinking, and design communication. She is part of a team of faculty innovators who originated Tandem (tandem.ai.umich.edu), a tool designed to help facilitate equitable and inclusive teamwork environments.

Gongjun Xu, University of Michigan

Dr. Gongjun Xu is an Associate Professor in the Department of Statistics with a joint appointment in the Department of Psychology at the University of Michigan.

Rebecca L Matz, University of Michigan

Becky Matz is a Research Scientist on the Research and Analytics team at the Center for Academic Innovation at the University of Michigan. She directs and supports research projects across Academic Innovation's portfolio of educational technologies and online learning experiences. Prior to joining Academic Innovation, she focused on STEM education assessment and research, connecting faculty with data, and developing interdisciplinary activities for introductory chemistry and biology courses at Michigan State University. Becky earned her Ph.D. in Chemistry and M.S. in Educational Studies from the University of Michigan.

Predicting Team Function Using Bayesian and Cognitive Diagnostic Modeling Approaches

Abstract

Team-based learning is commonly used in engineering introductory courses. As students of a team may be from vastly different backgrounds, academically and non-academically, it is important for faculty members to know what aid or hinder team success. The dataset that is used in this paper includes student personality inputs, self-and-peer-assessments of teamwork, and perceptions of teamwork outcomes. Using this information, we developed several Bayesian models that are able to predict if a team is working well. We also constructed and estimated Q-matrices which are crucial in explaining the relationship between latent traits and students' characteristics in cognitive diagnostic models. The prediction and diagnostic 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.

Introduction

Team-based learning (TBL) was first introduced in the 1980s to address problems that arose from large class settings [1], [2]. Although TBL was first implemented in business schools, team-based pedagogy can now be found across engineering, medical, and social sciences programs all around the world. Even though TBL provides students and instructors with many benefits, students do not always benefit equally from this learning method due to issues with free-riders or social loafing, work allocation, and communication, among others [3], [4]. For example, some students might feel the need for themselves to take on more interesting parts of a project, leaving the menial, boring, or repetitive work to other passive teammates. Some teams with mix-gendered teammates were found to have unequal work distribution with men doing more technical work, while women were doing more work related to communication or planning [3]. Thus, in order to ensure students are able to enjoy the benefits of TBL, teamwork assessment and support tools such as CATME or Tandem can be used to monitor the students' performances and notice any changes within the team [4]–[7].

By using teamwork assessment and support tools, learning analytics can be performed to optimize students' learning experiences. The large amount of data collected by the teamwork assessment and support tools provide an opportunity for researchers and instructors to detect various changes and relationships which are difficult to be detected in small samples. Furthermore, most teamwork assessment and support tools allow instructors and researchers to use student feedback to identify students or teams that require attention so that they are not struggling with academics, especially with the teamwork process. For example, with the help of teamwork assessment and support tools, 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 [4], [5]. Such benefits might not be realized if teamwork assessment and support tools are not used to collect a large amount of student data in class.

This study utilized data collected via Tandem, a teamwork assessment and support tool capable of providing formative feedback to teams and team members. In order to measure the

changes within the teams and check on students' progress, Tandem collects students' information through several different surveys. Tandem was first implemented in 2019 and has collected responses from more than 5000 students. In this paper, information from the "beginning-of-term" survey (BoT) and the weekly team check surveys (TC) was studied. These two surveys are described in the Data subsection of this paper. The aim of this study is to perform prediction and diagnostic analysis. Bayesian models are used to perform prediction in this paper while cognitive diagnostic models are used to perform diagnostic analysis. These models are described in detail in the Methods section.

Methods

Data

We focus on first-year engineering students at a large, public, semester-based, research-intensive university. The dataset consists of Tandem survey responses collected from students enrolled across 14 different sections of an introductory engineering design course between Spring 2020 and Fall 2021 (four semesters total). Owing to the COVID-19 pandemic, Fall 2020 and Spring 2021 courses were conducted online or in a hybrid mode. Nonetheless, team-based learning components were present in all courses. we used information from the BoT and the TC to predict whether the teams are working well weekly. The predicted results were compared to the students' actual weekly work well scores to determine the accuracy of the model. Due to the small amount of data present (65 students), we decided not to include the data of the students whose answer to the Gender question was not Male or Female.

Beginning-of-Term Survey

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 surveys 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, 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 answers choices from the BoT that were used in this study.

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.
[BT_Belongingness]	I expect to fit right into $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. Survey questions and answer choices asked in BoT. "$Course" is replaced by text describing the course (or sometimes, non-course context).

Team Check Survey

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 answer 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), 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.

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

Analysis

This paper focuses on two tasks: prediction and diagnosis. We used Python and Stan on Google Colab to build the Bayesian models and R in RStudio to build the Cognitive Diagnostic Models. The prediction section consists of developing three different Bayesian models to predict students' responses to the "Working Well" item, indicating how well the team is working together internally. The first two models were created with a new response variable that holds binary values. Thus, the models assume the likelihood function of the Bernoulli distribution. The

third model was created using the original response variable that holds 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. For each model, a probabilistic model with twenty covariates from the TC and the BoT was created to study how they affect students' work-well scores. The three Bayesian models developed were fitted and evaluated separately according to the steps in Box's Loop. Box's Loop is the iterative process of building a model based on data, using an inference algorithm to approximate the posterior, using the posterior to test the model against the data, and identifying the important ways that it succeeds and fails. If it fails, one will go back to the first step and build a new model [8].

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 [9], [10] 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 [11]. We also used our experience to come up with one expert-defined Q-matrix (Table 2) 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 [12].

Prediction: Bayesian Modeling

We are interested in studying the set of variables that are related to positive team work-well scores. Specifically, we want to find out the set of variables that can be used in a model to classify if a team is working well or not. The predicted results are compared to the students' actual responses to determine the accuracy of the models. The motive for using a Bayesian approach in predicting students' responses is that since the team support tool will continue to be used in future semesters, the Bayesian approach can ensure that past information about a parameter can be used to form a prior distribution for future analysis [13], [14].

We used three different Bayesian models in this paper. Both the Logistic and Hierarchical Logistic regression models have a Bernoulli distributed likelihood function as these models are predicting binary response variables. The Ordered Logistic regression model has an ordered logistic distributed likelihood function as it is used for predicting ordinal variables. Nonetheless, the predicted ordinal variables will be converted into binary variables to test for model accuracy since we are interested in learning whether the teams are working well or not. All three models were written using the Stan programming language with help from its documentation [15].

Logistic

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 [15, Ch. 15.2].

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

where

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

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

Hierarchical Logistic

A hierarchical logistic regression model was used as the data contain binary response variable 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.

$$\text{Cluster Index} = \text{ID}*2 + \text{Gender} - 1 \quad (2)$$

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).

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

where

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

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

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

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

Ordered Logistic

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.

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

where

$\eta = \beta^T x_i$

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

Diagnosis: 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 [16]. 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 [17] and between questions in surveys [18].

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 [11]. 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* [19]. According to [19], 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

$$P(X_j = 1|\alpha_{lj}^*) = P(\alpha_{lj}^*) \quad (5)$$

Although there are multiple link functions discussed in [19], this paper uses only the identity link function which is given in Equation (6).

$$P(\alpha_{lj}^{*}) = \beta_{j0} + \sum_{k=1}^{K_j^{*}} \beta_{jk}\alpha_{lk} + \sum_{k'=k+1}^{K_j^{*}} \sum_{k=1}^{K_j^{*}-1} \beta_{jkk'}\alpha_{lk}\alpha_{lk'}\ldots + \beta_{j12\ldots K_j^{*}} \prod_{k=1}^{K_j^{*}} \alpha_{lk} \qquad (6)$$

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 $\alpha_{k'}$;

$\beta_{j12\ldots K_j^{*}}$ is the interaction effect due to α_1,..., $\alpha_{K_j^{*}}$

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 [11], [20], 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 1.

Table 1. Q-matrix corresponds to three math questions and three latent attributes

Questions	Addition	Subtraction	Convert mixed number 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 [11]. In this paper, the Q-matrices were either estimated from the GDINA model or defined by experts. Workload, Confidence, and Sharing Idea are three latent traits of students related to what TC is characterizing. Table 2 shows the Q-matrix defined by the experts.

Table 2. 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
BT_Belongingness	1	1	1

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 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 (6). 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 (7) and (8). 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 $\alpha \in \{0, 1\}$, $1 \leq k \leq K$, and $\delta_{ji} = \alpha_{iK} \ldots \alpha_{i1}$ 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.

$$t_{jk} = \sum_{k=1}^{K} \delta_{ji} \text{ where } \alpha_{ik} = 1 \quad (7)$$

$$\widehat{Q}_{jk} = 1 \text{ iff } t_{jk} \neq 0 \quad (8)$$

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 (7), 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 (8), we get, $\widehat{Q} = (1, 1, 0)$. In (8), 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 (8). The estimated Q-matrix in (8) 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 is done using the orderQ function in cdmTools [21].

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 (7) and (8).
end

Results

Prediction: Bayesian Modeling

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 [22]. 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 [23]. 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 3, 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 the Logistic, Hierarchical Logistic, and Ordered Logistic 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) [24].

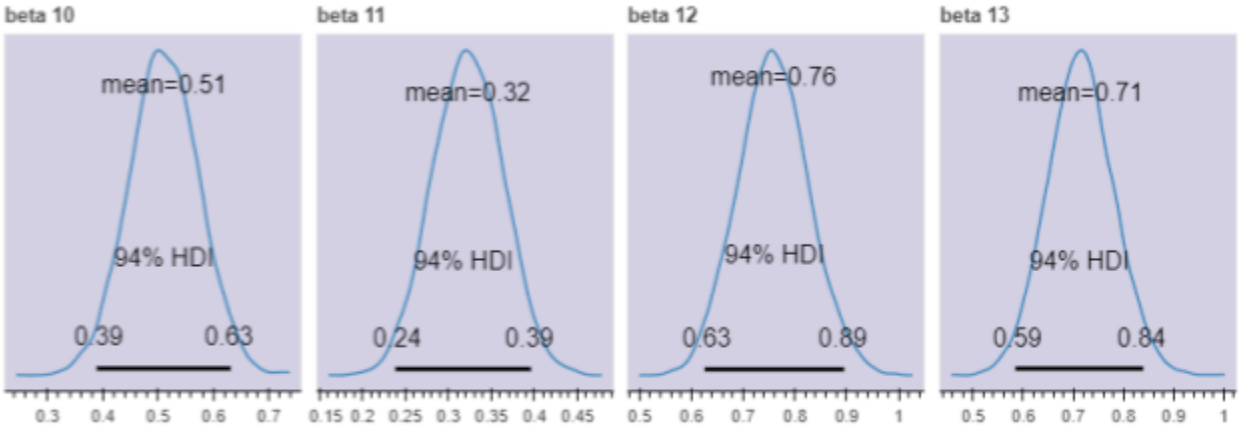

Fig. 3. A portion of the four posterior distributions of the beta ('TC_Workload', 'TC_Logistics', 'TC_Confidence', and 'TC_IdeaEquity') for the Logistics Regression model.

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.

Table 3. 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.

Diagnosis: Cognitive Diagnostic Modeling (CDM)

From the prediction section, 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 4 contains the results for the information criterion for the twenty-one estimated Q-matrices.

Using Algorithm 1, 20 different Q-matrices were estimated. In order to determine the Q-matrix that can best express the relationship between the J items and K latent skills, the 20 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 4, Model 3 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 [11], [25]. Both the Q-matrices 3 and Q0 are generically identifiable after checking with the identifiability conditions in Theorem 4 of [26, Sec. 5].

Table 4. Results for twenty-one Q-matrices.

Q-Matrix	AIC	BIC	Q-Matrix	AIC	BIC
1	7435.3	7676.4	12	7447.2	7697.7
2	7427.7	7621.6	13	7432.9	7655.1
3	7426.6	7629.9	14	7435.3	7676.4
4	7453.6	7647.4	15	7432.2	7692.2
5	7435.1	7714.0	16	7432.2	7692.2
6	7431.3	7615.7	17	7430.7	7671.8
7	7431.3	7615.7	18	7435.1	7714.0
8	7432.4	7597.9	19	7434.0	7694.0
9	7432.4	7597.9	20	7460.5	7720.6
10	7427.6	7687.6	Q0	7429.6	7576.2
11	7432.9	7655.1			

Table 5. Q-matrix 3 (left) and Q-matrix Q0 (right)

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

	Attr1	Attr2	Attr3
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

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

Conclusion 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.

There are some improvements that 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 diagnosis 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, or SDINA, or by using other estimation algorithms such as EM stepwise estimation with a provisional Q-matrix [11] or Restricted Boltzmann Machines [20].

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] 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.

[4] 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

[5] 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

[6] 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.

[7] 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

[8] D. M. Blei, “Build, Compute, Critique, Repeat: Data Analysis with Latent Variable Models,” *Annu. Rev. Stat. Its Appl.*, vol. 1, no. 1, pp. 203–232, Jan. 2014, doi: 10.1146/annurev-statistics-022513-115657.

[9] 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

[10] 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.

[11] 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.

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

[13] 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

[14] “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).

[15] S. D. Team, *Stan Functions Reference*. Accessed: Feb. 03, 2023. [Online]. Available: https://mc-stan.org/docs/functions-reference/index.html
[16] M. von Davier and Y.-S. Lee, Eds., *Handbook of Diagnostic Classification Models: Models and Model Extensions, Applications, Software Packages*. Cham: Springer International Publishing, 2019. doi: 10.1007/978-3-030-05584-4.
[17] Y. Liu *et al.*, "Collaborative Learning Team Formation: A Cognitive Modeling Perspective," in *Database Systems for Advanced Applications*, Cham, 2016, pp. 383–400. doi: 10.1007/978-3-319-32049-6_24.
[18] 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.
[19] 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.
[20] 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.
[21] 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
[22] "Stan Documentation," *stan-dev.github.io*. //mc-stan.org/users/documentation/ (accessed Dec. 11, 2022).
[23] "MCMC Sampling — CmdStanPy 0.9.64 documentation." https://cmdstanpy.readthedocs.io/en/stable-0.9.65/sample.html (accessed Feb. 07, 2023).
[24] "arviz.summary — ArviZ 0.11.2 documentation." https://oriolabril.github.io/arviz/api/generated/arviz.summary.html (accessed Feb. 04, 2023).
[25] 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.
[26] 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.