Graphically Modeling Student Knowledge: A Bayesian Network Approach

BY

XIAODAN TANG

B.A., Nanjing Normal University, 2012

M.A., University of Michigan – Ann Arbor, 2013

M.S., University of Illinois at Chicago, 2019

DISSERTATION

Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Educational Psychology
in the Graduate College of the
University of Illinois at Chicago, 2020

Chicago, Illinois

Defense Committee:

Yue Yin, UIC Department of Educational Psychology, Chair and Advisor

George Karabatsos, UIC Department of Educational Psychology

Yoon Soo Park, UIC College of Medicine

James Pellegrino, UIC Learning Sciences and Psychology

Hua-Hua Chang, College of Education, Purdue University

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor, Dr. Yue Yin, for her academic guidance, mental encouragement, endless patience throughout my doctoral study. Her wisdom continues to be a guiding light in my life and career.

I would like to sincerely thank Dr. George Karabatsos, for his knowledge in statistics, his vision in research, and his eye for details, all of which have broadened my research path.

I would also like to express my appreciation to each of my other committee members, Dr. Yoon Soo Park, Dr. James Pellegrino, and Dr. Hua-Hua Chang. They provided me with valuable feedback and insightful suggestions, which greatly help me improve my dissertation.

Finally, a big thank you to my family and my husband, whose tender loving care makes anything possible, and my friends, who have accompanied me through this joyful and exciting, but sometimes frustrating journey.

XT

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

CHAPTER PAGE

LIST OF TABLES

TABLE PAGE

LIST OF TABLES (Continued)

LIST OF FIGURES

FIGURE PAGE

LIST OF FIGURES (Continued)

LIST OF FIGURES (Continued)

FIGURE PAGE

LIST OF ABBREVIATIONS

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
BKT	Bayesian Knowledge Tracing
BN	Bayesian Network
CAF	Conceptual Assessment Framework
CAT	Computer Adaptive Testing
CDM	Cognitive Diagnostic Model
C-RUM	Compensatory Reparametrized Unified Model
DAG	Directed Acyclic Graph
DINA	Deterministic-Input, Noisy-And-gate
DINO	Deterministic-Input, Noisy-Or-gate
EAP	Expected A Posteriori
ECD	Evidence-Centered Design
EM	Expectation-Maximization
HMM	Hidden Markov Model
ICC	Item Characteristic Curve
IRF	Item Response Function
IRT	Item Response Theory
ITS	Intelligent Tutoring System
LD	Local Dependence

LIST OF ABBREVIATIONS (continued)

MCMC	Markov Chain Monte Carlo
MAP	Maximum A Posteriori probability
MIRT	Multidimensional Item Response Theory
MLE	Maximum Likelihood Estimation
NC-RRUM	Non-compensatory Reduced Reparametrized Unified Model
POKS	Partial Order Knowledge Structure
PPMC	Posterior Predictive Model Checking
PPP	Posterior Predictive Probability
RMSD	Root Mean Square Deviation
RMSE	Root Mean Square Error
SE	Standard Error
SD	Standard Deviation
TIMSS	Trends in International Mathematics and Science Study

SUMMARY

It is commonly believed that diagnostic information that are instructionally relevant and educationally meaningful would help students identify remediation learning paths and assist teachers customize their instruction according to students' knowledge gaps (DiBello et al., 2006). The Bayesian Network (BN) approach operationalizes cognitive diagnosis in a novel way of releasing some constraints of Cognitive Diagnostic Models (CDMs) and catering to a need of complex modeling of attribute structure and polytomous attributes. It also naturally relates to the assessment framework of Evidence-Centered Design (ECD). However, empirical studies fall short of systematically addressing the utility, uniqueness, and value of using BN in analyzing diagnostic assessment data.

In this study, I first conducted simulations to examine the performance of the BN approach across assessment scenarios of different sample sizes, test lengths, Q matrix complexities, and attribute types. Second, I evaluated the mastery classification accuracies of the BN approach when various amount of information on the structure in attributes is provided. Third, I compared the utility and the performance of both BN and CDM in terms of mastery classification accuracies across different assessment scenarios. Finally, I applied BN to analyze a dataset with dichotomous attributes from Trends in International Mathematics and Science Study (TIMSS) and a dataset with polytomous attributes.

SUMMARY (continued)

The results supported that BN can yield model parameters with acceptable accuracy for formative diagnostic assessments under various conditions, namely different test lengths, sample sizes, Q matrix complexities, and attribute types. BN can also provide adequate estimation results when partial information on the attribute structure is provided. The comparison of BN and a CDM model highlights the flexibility of BN in handling different assessment types. Finally, the real data analyses showcased the diagnostic reports on student performance levels based on the BN approach.

1. INTRODUCTION

1.1 Statement of the Problem

The question of how assessment can contribute to teaching and learning has long been of interest to researchers, practitioners, and educators. In his presidential address at the annual meeting of the National Council on Measurement in Education, Bennett (2018) presented his insights to the features of future educational assessments and emphasized the attempt to improve learning through “personalized” assessments, which can better accommodate diversity among students, and the incorporation of new approaches to modeling and analyzing assessment data. In the wake of the popularity of data science, researchers have paid attention to a combination of educational assessment with advances from data mining techniques in an effort to provide a more comprehensive understanding of students’ development of knowledge and skills, identify each individual student’s achievement gaps, and as a result aid improving instructors’ teaching and students’ learning. Bayesian Network (BN) is viewed as one of the well-developed machine learning and data mining techniques. It can operationalize the abovementioned goals of analyzing formative and diagnostic assessment driven by the cognitive theory of the measured subject domain and at the same time handling complex models (Almond et al., 2015). However, empirical studies are lacking in addressing the utility, uniqueness, and value of using BN in analyzing assessment data for a diagnostic purpose. This study intends to fill this gap.

1.2 Background

As educational agenda are inspired by the realization of students’ academic potential, educational researchers are motivated to explore ways of tapping the evidence of students’ academic achievement while promoting instruction and learning. One way to achieve this goal is developing assessment to measure students’ achievement and to make decisions on whether their performance meets the standards. Importantly however, it has long been a concern that the decisive and summative features of a typical assessment based on a pass-fail decision or a single ability estimate may be incongruent with the educational aim of guiding students to make improvement.

1.2.1 Cognitive Diagnosis

One long-standing approach to rectify the single ability estimate is to conduct a formative assessment, which is used to provide feedback for students and teachers in an ongoing learning process, so that they can improve both teaching and learning (Black & Wiliam, 1998, 2005). Further, Black and Wiliam (2005) contended that diagnostic score reports could serve as facilitators of formative assessment when initiated to achieve the three goals of diagnosing students' understanding of the measured subject domain, measuring the effectiveness of instructional strategies, and providing feedback to both teaching and learning. In concert with this claim, Yin et al. (2014) called for an application of formative assessment aligned with learning progression to facilitate conceptual change of domain knowledge. In a study of using the Rule Space Method (Tatsuoka, 1983) to diagnose students’ proficiencies in statistical knowledge, Im and Yin (2009) suggested that the diagnostic information obtained from the analyses would be helpful for students to fill their knowledge gaps. If a standardized test can provide diagnostic feedback regarding students’ attribute or sub-skill performance, it may hold pronounced promise for informing students’ knowledge gaps and prioritizing their remediation learning paths. A diagnostic performance report would be able to place students into various performance levels on each attribute to help them strategize their learning improvement process.

To illustrate this contention, we can consider a scenario where teachers and psychometricians collaborate to design a math test. The test can be designed to provide diagnostic, finer-grained feedback to learners about their mastery of the measured math skills and to inform them of the pathways to remedy those skills they have not yet sufficiently mastered. In this context, this math test is composed of the attributes of multiplication and addition. When the test is administered, and the test results are reported to students, it could be conceivable that two students with the exact same total score could have different understanding and competence regarding different attributes. For example, one student could have incorrectly answered all the addition questions and a second student all the multiplication questions, yet both receive the same total score. In order to differentiate the two students and build more informative score reporting, a diagnostic feedback mechanism could elaborate on students’ attribute performance given their overall performance and make a summative score report more useful than a total score for learners, their parents, and teachers. It may further motivate students to restructure and improve their knowledge and work toward achieving a better proficiency level.

The attribute score reports may offer comparisons among students with respect to their attribute performance, rankings by each attribute, and pass/fail information. From the teachers’ perspective, students’ attribute performance allows them to understand their weaknesses and strengths and then determine the best exercise strategy for each individual student or for students of the same attribute mastery profile. From the stakeholders’ perspective, students’ attribute performance could provide schools and test preparation centers with information to examine their curriculum readiness and learn potential remediation policies, so that they can further identify deficiencies in content standards, curricula, and instructional practices (Haberman et al., 2009). In other words, stakeholders expect tests to furnish an overall total test score, which can be used to place a student, or grant a certificate or license, while at the same time yield diagnostic information for student remediation, institutional evaluation, or program effectiveness (Feinberg & Wainer, 2014; Puhan et al., 2010).

Theoretically, the diagnostic feedback mechanism based on the observed students' performance should be delineated by the underlying processes embodied by assessment frameworks in order to aid its implementation and interpretation. Assessment Triangle (Pellegrino et al., 2001) and ECD (Mislevy et al., 2003) have been two widely adopted assessment frameworks, with the latter practically implementing the former (DiBello et al., 2006). The Assessment Triangle is comprised of three corners: the cognition component, the observation component, and the interpretation component. To situate assessment in a more practical sense, ECD brings together test developers and other stakeholders in order to define and support each piece of assessment development process. From the very beginning of identifying each target component of measurement, evidence through score interpretation is gathered to support decision-making (Huff et al., 2010). Given that ECD plays an important role in reporting attribute performance, measurement models can coordinate with ECD in providing diagnostic information. BN and Cognitive Diagnostic Models (CDMs) are well known as classification approaches and can be conceptually relating to ECD.

1.2.2 CDMs and BN

CDMs can be used to demonstrate how well assessments classify students' level of proficiency. Learners and teachers can get additional information from formative assessments, if the assessments are designed under the Assessment Triangle and the ECD frameworks, and subsequently calibrated with CDMs for their assessment context. Lee et al. (2011) found that the detailed diagnostic feedback provided by CDM would directly contribute to classroom instruction. Later, Park and Lee (2014) reparametrized one of CDMs to account for the impact of covariates on students' mastery of the measured latent skills. This finding yields insights into ways of enriching the complexity of CDM to illuminate students' learning process. A recent study conducted by Tu et al. (2019) considers the attribute hierarchy when using CDMs and suggested that the classification results might be misleading if ignoring the hierarchy.

Despite of the advances of CDMs over other measurement models, CDMs were not widely applied to analyze assessment data due to some constraints. Culbertson (2016) pointed out that CDMs fail to specify the structure of the underlying latent abilities. Further, Rupp et al. (2010) explained specific and constrained attribute mastery status for a correct response in each CDM model. To rectify these constraints, BN, which can construct a joint probability distribution over the involved variables based on graphical representations, serves as one novel way to build a flexible modeling framework. Its modeling construction can rely on assessment frameworks, and its modeling results can infer diagnostic performance report (Almond et al., 2015; Rupp et al., 2010). It also has the capacity of identifying the hierarchy in attributes by deriving a series of conditional probabilities based on Bayes' Theorem and releasing the condensation rule in CDMs by specifying probabilities of a successful response for each possible mastery profile.

The application of Bayesian inference research has drawn attention in the field of modeling assessment data. For example, Karabatsos and Sheu (2004) proposed an order-constrained Bayes inference framework to analyze dichotomous item responses. Additionally, software functions and packages pertinent to Bayesian inference have also been developed for analyzing educational test data (for example, software of Karabatsos, 2017). Inspired by the BN framework, Mislevy et al. (1999) proposed a Bayesian framework for modeling educational assessment that considers a joint distribution over responses, students, and items by breaking a holistic structural model into small models. Methodologically, BN can capture the conditionally probabilistic relationships among latent variables and observed variables, based on student responses to individual assessment tasks or test items and latent skills of interest indicated by theory. Using BN to propagate information from assessment is consistent with the role of assessment that can tap students' performance and make inferences on students' learning process in a particular subject domain (Almond et al., 2015). As such, this recognized connection can be manifested by constructing a structural graph of the measured component skills, an advance over CDMs.

Like CDMs, BN builds on a theory-based structure which incorporates content experts' input about the dependent relationships among attributes. However, different from the entire reliance on expert input or theory-based structure in CDM, BN can update these prior beliefs based on observations. Additionally, BN can release the compensation and condensation rules of the attribute mastery pattern of CDMs by probing each combination of attributes in a successful response. Another advantage of BN is the flexibility in modeling complex and evidentiary relationships among attributes and items.

A BN analysis is reminiscent of but not identical to the CDM analysis, providing diagnostic feedback to students based on their test performance (Rupp et al., 2010; Zhang & Chang, 2016). Both frameworks are extended versions of latent class models. Further, CDMs can extend IRT modeling while BN can be parameterized using IRT. Despite the importance of diagnostic feedback (De La Torre & Minchen, 2014), the majority of research on cognitive diagnosis has focused on the application of CDMs rather than on BN (Shute, 2011; Xenos, 2004). Although both approaches aim to provide informative attribute feedback regarding student performance, few studies on educational assessment compared CDMs and BN. Therefore, their similarities and differences in terms of utility and values are not well understood. To be scaled up for further research, it highlights a practical need to examine how BN modeling performs in analyzing student overall ability and their attribute performance, and what diagnostic feedback BN could help to provide that would further promote student learning, and finally whether it could furnish more information regarding students' learning strategies relative to CDMs.

Taken together, this study concerns measurement model, statistical inference, and cognitive diagnosis in the analysis of assessment results using a simulation and a data application of the BN approach. Theoretically, the ECD approach holds the promise for integrating cognitive task design and statistical measurement modeling and constructing a framework that informs the design of an assessment and the interpretation of assessment results (Yan et al., 2003). Further, it builds blocks for a cognitive diagnosis of students' domain knowledge. In the measurement and statistical sense, BN dismantles the underlying process of students' learning and usage of knowledge or skills to respond to items, and further provides a more comprehensive guidance of future learning paths rather than a pass-fail or a single ability estimate, which may convey less informative evaluation. Below I review the literature to build the foundation for my study.

2. REVIEW OF THE LITERATURE

2.1 Organization of the Literature Review

In this chapter, I address the assessment frameworks, the measurement models, and the connection between them. First, I introduce the assessment frameworks of Assessment Triangle and ECD as the underlying foundations for assessment data analysis. Second, I discuss about the main principles of the measurement models, including IRT, CDMs, and BN. Third, I review the psychometric studies on the application of the BN method and unfold the common evaluation criteria to determine the accuracy and precision of model results of simulation studies. Finally, I map the components of ECD to the measurement models so that the three overarching topics of cognitive domain diagnosis, measurement modeling, statistical inference are integrated.

2.2 Assessment Triangle and Evidence-Centered Design

The framework of Assessment Triangle is proposed by Pellegrino et al. (2001). It is composed of three vertices: cognition, observation, and interpretation (see in **Error! Reference source not found.**) that must work congruently in assessment development. Cognition reveals how student proficiency of the latent trait is presented and developed. Observation refers to the tasks or scenarios used to tap student performance as evidence of the measured latent traits. Further, the observation activities are represented by the design and specifications of assessment tasks. Interpretation illuminates the connection between observation and cognition. In other words, the observation of students' representation of their knowledge is displayed through interpretation. Practically speaking, the interpretation helps researchers to determine an appropriate measurement model. The assessment triangle provides a valuable theoretical framework to develop and use assessments.

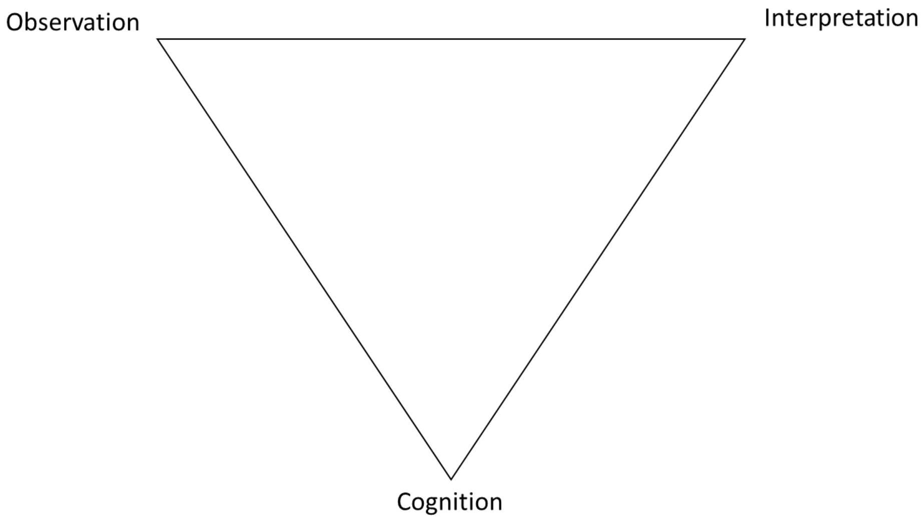


Figure 1. Assessment triangle.

ECD is an assessment design framework that originates from the Assessment Triangle. It also requires evidentiary arguments during assessment design process. It serves as a tool to systematize the design and development of assessments based on a logical connection between and within assessment goals, the cognitive framework of domain knowledge, the design of tasks, the collection of performance evidence, and an operational delivery of assessments. As has been demonstrated before, based on the theoretical framework of the Assessment Triangle, ECD provides a template for assessment design and analysis guided by evidentiary reasoning and knowledge representations.

The ECD framework has five layers that spread from the knowledge presentation of domain analysis, modeling, to a conceptual assessment framework (CAF) essential to test design, and finally to the test operationality of assessment implementation and delivery. Mislevy and Haertel (2006) demonstrated details on ECD layers. Domain analysis is the first step of assessment design. For this step, test development team needs to collect all kinds of information on the measured domain including concepts, tools of knowledge representations, and conventional application formats. Domain modeling is a narrative model of logical relationships among the components concluded from the domain analysis. Researchers and test designers construct domain modeling to guide and organize the descriptive statement of the CAF.

The CAF details the design blueprint of an assessment by incorporating models for assessment arguments based on the first stage of domain analysis and modeling. It builds a bridge between assessment arguments and assessment activities. CAF specifies the proficiency model with knowledge, skills, and abilities involved in the assessment; the evidence model demonstrating the student acquisition of knowledge and skills; the task model describing the features of tasks to elicit the expected evidence of student achievement.

To be specific, the proficiency model represents the relationship of the measured components. It lays out the inferences made about student proficiencies. The variables in the proficiency model can be probed from the nature of students' performances and their reasoning process during a real assessment through the lens of ECD.

The evidence model, explaining how we measure the domain knowledge, consists of two sections: the evidence rules component characterizing the evidence identification process and the measurement model summarizing evidence. The evidence rules component considers the quality of expected student responses and their accuracy, strategies they use for problem-solving, etc. Item responses capture student performance, and then teachers can evaluate their item responses based on predetermined scoring rules to yield observations of student performance. The evidence rules component provides information about proficiency model and feedback on student performance. As such, the evidence rules model is represented by item scoring and scoring results. Teachers or researchers use answer keys or scoring rubric to evaluate how well students have used the domain knowledge according to the relationships specified in the domain modeling. Additionally, guided by the evidence rules model, test designers need to concern whether tasks assess enough information to evaluate the quality of student work.

After the scoring procedure, the measurement model is to define and quantify the degree to which item responses reveal student performance. It provides information on the connection between the proficiency model variables and the observed item responses based on statistical models that can be applicable to assessment data. Psychometric models play an important role in the measurement framework, including the widely used classical test theory and IRT, and the less familiar CDMs and BN. They accumulate and synthesize evidence on student proficiency in the targeted domain by collecting evidence from the proficiency model and the evidence rules procedure.

For the questions of where to measure the domain knowledge, the task model describes the tools where students can produce responses and provides information on their proficiency. Also, teachers or researchers design tasks or items and describe test specifications guided by the task model. The test specification includes test prompts presented to students and test items for collecting student responses. Test designers need to make decisions on the formats of items and the features of these tasks that are salient to student proficiency.

The assembly model specifies the breadth and diversity of the domain knowledge as assessment constructs. The assembly model concerns the relationship among the proficiency model, the evidence model, and the task model to form a complete assessment tool. The rules to construct a form depend on validity, which is referred to how accurately each student variable can be measured, and balance, which decides the distribution of content knowledge in an assessment to maintain an appropriate breadth and diversity.

The presentation model describes the style on how task materials should be presented in terms of test formats, platforms, and any other alternatives. The delivery system model encompasses a big picture of all assessment pieces that convey evidentiary arguments from the collection of proficiency, evidence, task, assembly, and presentation models. Tests for different purposes may have different models. For example, if we want to use the same test for both diagnostic purposes and placement purposes, the task models might be the same but evidence models and proficiency models will be different. The proficiency model of a diagnostic assessment will include examinees' proficiency levels on each attribute, while that of a placement assessment will rely on examinees' overall competence level. Accordingly, the evidence model of a diagnostic assessment will list the connections of each item and each attribute, while in the placement assessment, items may only vary by item-level parameters (e.g., item difficulty).

2.3 Measurement Models

Researchers and practitioners use measurement models to analyze educational or psychological data and make inferences on participants' competence of the latent trait of interest.

One of the typical measurement models is IRT, which has been well developed and widely adopted in a substantial body of literature. Typically, IRT considers participants' achievement and item performance in a general level and are applicable to various assessment purposes and formats. Additionally, different functions can be derived to serve for unidimensional or multidimensional situation, dichotomously or polytomously score items. Another measurement model is CDM, which aims for classifying students into different proficiency levels on each attribute of the latent skill. It has recently drawn great attention due to its diagnostic feature, which would provide informative feedback beneficial to teaching and learning. In the era of big data driven by data science, researchers have started to examine whether machine learning and data mining techniques can be effectively used to analyze assessment data. BN, as a technique of machine learning, has been applied into the measurement framework and the assessment data analysis. This section first briefly introduces IRT followed by CDM, and then elaborates on the principles of BN and its application in psychometrics.

2.3.1 Item Response Theory

Item response theory (IRT) is a theoretical framework for analyzing design and scoring of educational tests and psychological instruments (e.g., Embretson & Reise, 2000). These tests or instruments are established to measure latent variables that are not directly observable. As such, IRT is also known as latent trait theory, which describes the relationship between individual performance and the performance of the entire group on an overall measure of the latent construct. IRT considers items with different levels of difficulty rather than equal difficulty and also scales item difficulty and person ability onto the same metric and provides sample- and test-independent information on item characteristics and person performance (Embretson & Reise, 2000).

Mathematically, in IRT modeling, the probability of a correct response to an item is assumed to be a function specifying the relationships between person and item parameters. Each item is characterized by item parameters (e.g., item difficulty, item discrimination, and guessing). Each examinee is characterized by a person ability estimate. As such, in the IRT modeling, examinees' performance can be represented by latent traits (i.e., person abilities) estimated by their item responses. By convention, θ represents person ability, and it is usually assumed to be normally distributed. b, a, c respectively represents item difficulty, item discrimination, and guessing parameters. Based on this relationship, IRT provides item characteristic curve (ICC) for each item showing a bell-shaped probability curve across person abilities.

ICC reflects the relationship between ability and item performance, which is monotonically increasing. In other words, the probability of answering an item correctly always increases as ability increases. Students with lower level of proficiency would be estimated to have lower probability of getting the item correct, and vice versa. The ICC is depicted according to the mathematical function as shown in Equation (1) for the 1-PL IRT model or also called as the Rasch model. The 2-PL model (Equation 2) has two item parameters including item difficulty and item discrimination. The item discrimination parameter determines the slope of the ICC at the inflection point. Items with a steep slope are highly discriminating (i.e., a higher value of parameter a), while items with a gradual slope are poorly discriminating. Items with a high discrimination distinguish students with different levels of performance. As such, they are informative. When item discrimination equals 1, the function will be the same as 1-PL model. The third parameter in the 3-PL IRT model (Equation 3) is the guessing parameter c, considering that students may give a correct answer due to guessing. The guessing parameter is denoted as the lower asymptote of the ICC.

$$P_{\mathrm{j}}(\theta) = \frac{\mathrm{e}^{(\theta - b_{\mathrm{j}})}}{1 + \mathrm{e}^{(\theta - b_{\mathrm{j}})}}, \tag{1}$$

where *j* =1,2,3,...,*J* for items,

θ represents person ability,

b_{j} represents the item difficulty of item *j*,

$P_{\mathrm{j}}(\theta)$ represents the probability of getting item *j* correct.

$$P_{\mathrm{j}}(\theta) = \frac{\mathrm{e}^{a_{\mathrm{j}}(\theta - b_{\mathrm{j}})}}{1 + \mathrm{e}^{a_{\mathrm{j}}(\theta - b_{\mathrm{j}})}}, \tag{2}$$

where *j* =1,2,3,...,*J*,

θ represents person ability,

a_{j} represents the item discrimination of item *j*,

b_{j} represents the item difficulty of item *j*,

$P_j(\theta)$ represents the probability of getting item *j* correct.

$$P_{\mathrm{j}}(\theta) = c_{\mathrm{j}} + (1 - c_{\mathrm{j}}) \frac{\mathrm{e}^{a_{\mathrm{j}}(\theta - b_{\mathrm{j}})}}{1 + \mathrm{e}^{a_{\mathrm{j}}(\theta - b_{\mathrm{j}})}}, \tag{3}$$

where *j* =1,2,3,...,*J*,

θ represents person ability,

a_{j} represents the item discrimination of item *j*,

b_{j} represents the item difficulty of item *j*,

c_{j} represents the pseudo-chance ("guessing") parameter of item *j*,

$P_{\mathrm{j}}(\theta)$ represents the probability of getting item *j* correct.

IRT modeling relies on three assumptions: (a) The assumption of unidimensionality states that only one latent trait is measured by the test. If a test is designed to measure multiple latent traits, multidimensional IRT (MIRT) should be used for analysis. (b) The assumption of local independence states that each examinee's responses to a given set of test items are independent when conditional on the model parameters. If the IRT model parameters explain all the variances of item responses, then each pair of inter-item correlations becomes zero. In contrast, the violation of local independence, termed as local dependence, occurs when this conditional independence property fails to hold on a given set of item responses. The non-zero inter-item correlations exist among pairs of test items, even after conditioning on the IRT model parameters. Consequently, locally dependent items carry less information than the IRT model would predict. The assumptions of unidimensionality and local item independence are closely related given the fact that local item independence infers the unidimensionality of a test. (c) The third assumption states that the responses of a person can be depicted by the Item Response Function (IRF), which is defined by Equation (1), (2), (3) for 1-PL, 2-PL, 3-PL IRT models, respectively. The IRF of each model defines the shape of the ICC. After all the model assumptions are met, IRT would be able to yield invariant person parameters, suggesting that person ability remains the same across different tests, and invariant item parameters, suggesting that they also remain the same across examinees of different performance levels.

2.3.2 Cognitive Diagnostic Modeling

CDMs, as sophisticated measurement models, are intended to improve the quality of diagnostic feedback provided to students at various performance levels. Rather than assigning to examinees a single ability estimate on a continuous scale, CDMs aim to provide examinees with finer-grained information pertaining whether or not they have mastered each attribute required to answer a certain item correctly (Rupp et al., 2010). In this way, CDMs can be used to classify students into different mastery groups and further provide diagnostic feedback on their strengths and weaknesses of attribute knowledge. In this section, I introduce CDMs and their basic characteristics.

Given the predetermined finer-grained proficiency dimensions (e.g., attributes), CDMs are designed to furnish criterion-referenced interpretations for each attribute specific to each student or students within the same mastery group. In decision-making situations (e.g., placement, admission, or certification), students are usually classified into non-mastery or mastery on a specific domain. To evaluate students' learning, CDM can also be used to yield informative feedback based on more than two levels of classifications. For example, Templin (2004) has explored ways of assigning several mastery levels for each attribute. Moreover, similar as BN modeling, CDM is also capable of considering the structure in attributes (Leighton et al., 2004; Templin & Bradshaw, 2014), although this feature has not been fully explored and applied. In the following sections, I introduce the critical components of CDMs, Q matrix, and three commonly used CDMs.

Q matrix. Like other latent class models, the CDM classification results rely on how latent skills are measured across items. In other words, we need to input the information on which attributes are measured by each item to enable CDM to classify students' performance levels based on their responses to the items measuring the corresponding attribute. This input is a loading table called Q matrix (Tatsuoka, 1983). The specification of Q matrix is developed or confirmed by subject-matter experts. From an assessment perspective, Q matrix indicates how the measurement of each attribute is distributed across items. From a statistical perspective, Q matrix is a loading structure, like those often used in a confirmatory factor analysis, specifying the loadings of each attribute on items. Q matrix not only can help with the analysis of test data, but also assist in developing and revising assessment, as it clearly indicates the distribution of attributes across items.

The structure of Q matrix is composed of rows representing items and columns representing the measured attributes. A simple structure Q matrix is specified when each item is only dependent on one attribute. For a complex structure Q matrix, items can be loaded on multiple attributes. For a dichotomous classification of mastery and non-mastery on students' attribute performance, the Q matrix will consist of 1s and 0s indicating whether an attribute is measured by an item. For a polytomous classification of attributes, 0s still represent an attribute is not measured while the other non-zero integers represent different levels of mastery. An example of a dichotomous Q-matrix with binary entries and a polytomous Q-matrix with 3-level categorical entries are shown in Table I and Table II, respectively. In Table I, a test contains J = 3 items and measures K = 3 attributes. Item 1 is measured by attribute 1, item 2 is measured by attribute 2 and 3, and item 3 is measured by attribute 3. Table II shows the Q matrix for a test with polytomous attributes. In this example, items 1 requires high mastery of attribute 1 but neither attributes 2 and 3; item 2 requires high mastery of attribute 1 and medium mastery of attributes 2 and 3; item 3 requires medium mastery of attribute 3.

TABLE I AN EXAMPLE OF A SIMPLE STRUCTURE DICHOTOMOUS Q MATRIX

Items	Attribute 1	Attribute 2	Attribute 3
1.	1	0	0
2.	0	1	1
3.	0	0	1

TABLE II AN EXAMPLE OF A COMPLEX STRUCTURE POLYTOMOUS Q MATRIX

Items	Attribute 1	Attribute 2	Attribute 3
1.	2	0	0
2.	2	1	1
3.	0	0	1

A mathematical explanation of CDM is specified as follows. I used a *K*-dimensional vector: $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_K)$ to represent the collection of *K* attributes measured by a test. The items are denoted by $j = 1,\ldots,J$, where *J* is the test length of the test. Q matrix is used to describe the relationships between items and attributes, in which 1 indicates that the attribute is measured by the item, and *0* suggests that the attribute is not measured by the item. The Q matrix has *J* rows: each row refers to one item and provides information on which attributes need to be mastered in order to answer this item correctly. Examinees are denoted by $i = 1, \ldots, N$, where *N* is the total number of examinees. The response of examinee *i* to item *j* is denoted by X_{ij}. If the examinee correctly answers a dichotomous item, then $X_{ij} = 1$, otherwise $X_{ij} = 0$. The responses for all examinees to all items are denoted by **X,** which is an $N\times J$ matrix.

As discussed above, a complex structure of Q matrix may have multiple attributes load on one item. This feature leads to a discussion on how to specify the individual contribution of each attribute toward the overall problem-solving process of an item. Mathematically speaking, the combination of attributes can be either additive or multiplicative. According to this relationship, CDMs can be classified into two types of models: the compensatory model and the non-compensatory model. The non-compensatory model depends on the assumption that a successful response requires the mastery of all the measured attributes. In other words, a low value on one latent variable cannot be compensated by a high value on another latent variable. In this case, the non-compensatory CDM model will use a product to represent the relationships among attributes. The typical non-compensatory models in CDMs are the deterministic-input, noisy-and-gate model (DINA) and the non-compensatory reduced reparametrized unified model (NC-RRUM). In the generic compensatory models, the average or sum of the required attributes would influence the possibility of item responses, therefore a low value on one latent variable can be compensated for by a high value on another latent variable. The typical compensatory models in CDMs are the deterministic-input, noisy-or-gate model (DINO) and the compensatory reparametrized unified model (C-RUM). Below I illustrate the DINA model as an example of non-compensatory models, the DINO model as an example of compensatory models, and the generalized DINA (G-DINA) model as a saturated model free from the compensatory and non-compensatory rule.

The DINA model. The DINA model (Haertel, 1989; Junker & Sijtsma, 2001) is one of the simplest and most widely used CDM. As one of the non-compensatory models, it is assumed that examinees must possess all the measured skills to successfully answer an item. By doing so, students will be classified into two exclusive groups for each item: mastering all the required attributes and not mastering at least one of the required attributes. Although students in the latter group may have different mastery profiles (e.g., mastering one out of the three required attributes vs. mastering two out of the three), the DINA model will compute a same probability of correctly answering this item for the latter group.

Tatsuoka (1983) proposed the rule space methodology as one type of cognitive diagnosis, which lays the foundation for current CDMs. In terms of the model specification, Tatsuoka (1995) proposed the ideal response should be specified as follows:

$$\eta_{ij} = \prod_{k=1}^{K} \theta_{ij}^{q_{jk}} \qquad (4)$$

where

η_{ij} =1 when examinee *i* has mastered all the required attributes for item *j*,

K represents the total number of attributes measured by the test,

θ_{ij} represents the mastery status of person *i* on attributes required by item *j*,

q_{jk} represents the *k*th attribute required to solve the *j*th item.

The slipping parameter is shown in Equation (5), and the guessing parameter is shown in Equation (6).

$$s_{\mathrm{j}} = p\left(X_{\mathrm{ij}} = 0 \middle| \eta_{\mathrm{ij}} = 1\right) \quad (5)$$

where

η_{ij}=1 represents examinee *i* has mastered all the required attributes for item *j*,

$X_{\mathrm{ij}} = 0$ represents examinee *i* incorrectly answers item *j*.

$$g_{\mathrm{j}} = p(X_{\mathrm{ij}} = 1|\eta_{\mathrm{ij}} = 0) \quad (6)$$

where η_{ij}=1 represents examinee *i* has mastered all the required attributes for item *j*,

$X_{\mathrm{ij}} = 1$ represents examinee *i* correctly answers item *j*.

The formula for the slipping parameter represents the probability of a student responding to the item incorrectly while mastering all the measured attributes. The formula for the guessing parameter denotes that the probability of correctly answering the item while having not mastered at least one attribute required for the item. Given the ideal response, the item response function for the DINA model is given by Equation (7).

$$p(X_{\mathrm{ij}} = 1|\eta_{\mathrm{ij}}) = (1 - s_{\mathrm{j}})^{\eta_{\mathrm{ij}}} g_{\mathrm{j}}^{1-\eta_{\mathrm{ij}}} \quad (7)$$

where η_{ij} represents the mastery status of examinee *i* on all the attributes required by item *j*,

$X_{\mathrm{ij}} = 1$ represents examinee *i* correctly answers item *j*,

s_{j} represents the slipping parameter of item *j*,

g_{j} represents the guessing parameter of item *j*.

The DINO model. The deterministic input, noisy-**or**-gate (DINO) model **(e.g., Templin & Henson, 2006)** is compensatory analog to the DINA model (Rupp et al., 2010). Similarly, the DINO model is also parameterized by slipping and guessing parameters. But their ideal response function is different, shown as follows:

$$\eta_{\mathrm{ij}} = 1 - \prod_{\mathrm{k}=1}^{\mathrm{K}} \left(1 - \theta_{\mathrm{ij}}^{\mathrm{q_{jk}}}\right) \qquad (8)$$

where

η_{ij} = 1 represents examinee *i* has mastered at least one of the required attributes for item *j*,

K represents the total number of attributes measured by the test,

θ_{ij} represents the mastery status of person *i* on attributes required by item *j*,

q_{jk} represents the *k*th attribute required to solve the *j*th item.

As shown in the Equation (8), the ideal response function is defined by the situation where examinees master at least one of the required attributes to correctly answer the item. In other words, the DINO model assumes that examinees are likely to answer items correctly by mastering at least one of the required attributes rather than mastering all the required attributes. Although the formulas for the slipping and guessing parameters are the same as the DINA model, they have different interpretations. The slipping parameter formula (see in Equation 5) represents the probability of a student responding the item incorrectly when at least one measured attribute is present. The guessing parameter formula (see in Equation 6) denotes the probability of correctly answering the item while all measured attributes are absent. It is also the same case for the response function.

The GDINA model. The generalized diagnostic input noisy "and-gate" (GDINA) framework articulated by De La Torre (2011). Unlike specific CDMs (e.g., DINA, DINO) constrained by the compensatory rule, the GDINA model serves as a saturated model releasing this constraint and provide probabilities of success for students of each mastery profile. Specifically, the DINA model partitions students' mastery profiles into two classes: mastery of the required attributes and non-mastery of the required attributes. However, the G-DINA model partitions the latent classes into 2^{k} groups, which each has its own probability of success. The mathematical function of the G-DINA model can be specified as:

$$p(X_{\mathrm{j}} = 1|\theta_{\mathrm{lj}}^{*}) = \delta_{\mathrm{j0}} + \sum_{\mathrm{k=1}}^{\mathrm{K_j^*}} \delta_{\mathrm{jk}}\theta_{\mathrm{ljk}} + \sum_{\mathrm{k=1}}^{\mathrm{K_j^*-1}} \sum_{\mathrm{k'=k+1}}^{\mathrm{K_j^*}} \delta_{\mathrm{jkk'}}\theta_{\mathrm{lk}}\,\theta_{\mathrm{lk'}} + \cdots + \delta_{\mathrm{j12\ldots K_j'}} \prod_{\mathrm{k=1}}^{\mathrm{K_j^*}} \theta_{\mathrm{lk}} \qquad (9)$$

where *j* represent items, *k* represent attributes, θ denotes the attribute parameters, θ_{lj}^{*} denotes students' mastery status on the required attributes, δ denotes item parameters.

2.3.3 Bayesian Network

CDMs have been well researched in the literature but rarely been implemented in the operational context (Rupp et al., 2010). The reason might be that most CDMs are restrictive with constraints of assuming the attribute pattern of giving a correct response (i.e., compensatory rules). Consequently, when they disagree with the compensatory rule, teachers or researchers might have less passion in applying CDMs due to these constraints. For the same purpose of classifying students according to their ability on latent traits, BN, as a less restrictive modeling framework, has recently been applied as a method of cognitive diagnosis. Different from the dichotomous classification of mastery and non-mastery estimated by traditional CDMs, the levels of latent skill performance in BN are typically polytomous, indicating the degree of mastery on each attribute. Due to its flexibility in statistically modeling the relationship between observed and latent variables, BN has also been widely applied in the field of data mining and machine learning. In particular, the method of data mining is a process to discover patterns in large data sets, and machine learning is a process to learn from data and make predictions based on new data. Both methods have been applied into education research (Romero et al., 2010). This integration contributes to making inferences on educational outcomes in a diverse way and solving research questions that are not feasible with small sample, which is common in the educational setting.

Typically, the BN technique constructs a model from known information including prior information of knowledge structure from experts or the distribution we wish to model. The goal is to obtain and optimize a model that can more precisely capture the real distribution representing the observed data. BN uses probabilistic graphical models to account for conditional distributions over several variables (Koller et al., 2009). BN graphical models consist of nodes representing variables and edges (i.e., arrows) representing directed probabilistic relationships between variables as illustrated in Figure 2. The graph shows a model representing the problem involved with variables and parameters. In particular, BN is one type of graphical models that use directed arcs to construct a directed acyclic graph (DAG): directed means the edges are directed by arrows, and acyclic means that the graph has no cycles so that you cannot reverse the edges.

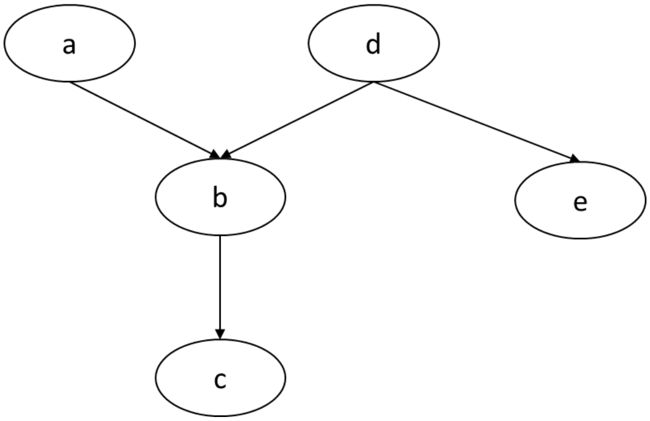


Figure 2. An example BN graph.

Representing in symbols, the graph of a BN model can be simplified by using *G* = (*N*, *E*). It consists of a set *N* of nodes along with a set *E* of directed edges. In Figure 2, an edge (*a*, *b*) denotes an arrow pointing from node *a* to node *b*: *a* is a parent of *b*; *b* is a child of *a*. Similarly, there are two arrows pointing to *b* from *a* and *d*, which indicates *a* and *d* are parents of *b*, and *b* is a child of both *a* and *d*. *d* is also pointing to *e*, suggesting another child of *d*. To explain the relationships among variables in a graphical model, each variable is represented conditionally on its parents, the set of variables with a directed edge into the variable. The general form of a dependent relationship for each variable is written in Equation (9):

$$p(\boldsymbol{X}) = \prod p(x|parents(x)) \qquad (10)$$

The values of nodes can be discrete or continuous. For example, Figure 3 shows a graphical structure of the dependent relationships among attributes for a sample test with two items measuring three attributes. In the example, attribute 1 and 2 are assumed to be independent. Attribute 3 depends on these two attributes. In the corresponding Q matrix (as shown in Table III), item 1 is loaded on attribute 3, and item 2 is loaded on attribute 2. It is a joint probability distribution that can be decomposed into smaller local probability relationships between the latent variables and the observed outcomes.

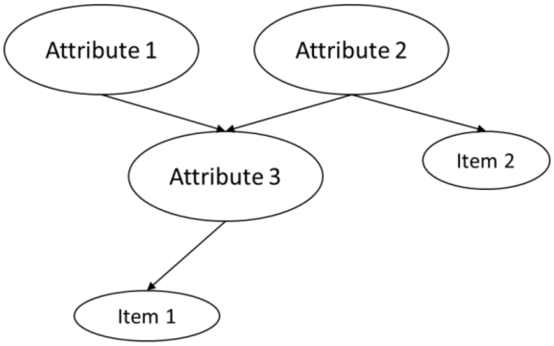


Figure 3. A DAG representing part of an assessment.

TABLE III THE Q MATRIX FOR THE GRAPHICAL EXAMPLE IN FIGURE 3

Items	Attributes		
	Attribute 1	Attribute 2	Attribute 3
Item 1	1	0	0
Item 2	0	1	0

Specifically, the conditional dependence distributed over all the variables is derived by a set of local probability distributions that represent the dependence of each variable on its parents. According to the chain rule, this derivation can be defined by a product of local distributions of each variable. For example, Equation 10 shows the factorization of the joint distribution associated with the DAG in Figure 3. p(A1), represents the distribution of students with different levels of mastery for attribute 1. p(A2), represents the distribution of students with different levels of mastery for attribute 2. The distribution over attribute 3 is a conditional distribution, p(A3 | A1,A2), specifying the distribution over attribute 3 is dependent on attribute 1 and attribute 2. The probability of student mastery of attribute 3 depends on their mastery of attribute 1 and attribute 2. As such, each student would have their own distribution of attribute 3 given their proficiency in attribute 1 and 2.

$$p(\mathrm{A1}, \mathrm{A2}, \mathrm{A3}, \mathrm{I1}, \mathrm{I2}) = p(\mathrm{A1})p(\mathrm{A2})p(\mathrm{A3}|\mathrm{A1}, \mathrm{A2})p(\mathrm{I1}|\mathrm{A3})p(\mathrm{I2}|\mathrm{A2}) \quad (11)$$

As a concrete example, we might assume that a student who have fully mastered both attribute 1 and 2 has 70% probability of fully mastering attribute 3, 20% partially mastering attribute 3, 10% not mastering attribute 3. Conversely, a student who have fully mastered attribute 1 but partially mastered attribute 2 may only be 50% probable to fully master attribute 3. In general, each variable in the graphical model has a conditional probability distribution (CPD), depending on the joint distribution of its parents in the model. If a variable has no parents, its CPD appears to be its own probability distribution with different magnitude of its values. To construct prior conditional relationship among variables, we may depict the edges based on our assumptions or content experts' suggestions.

From the modeling framework and mathematical specification, we may find some similarities between the two approaches. CDMs are developed to classify students into different proficiency levels, and BN can be used to serve the same purpose. Other models can do similar cognitive diagnosis, but they are not very comparable to either CDM or BN. This highlights the need to compare and evaluate the performance of the two approaches.

2.3.4 Bayesian Network and Measurement Models

As discussed before, BN has become a useful tool to understand the relationships among a large number of variables. Many studies have explored its promising application in analyzing educational or psychological data relative to other measurement models. BN highlights its advantages in a powerful modeling flexibility for building psychometric models and supporting a wide range of assessment types and contexts, particularly for multidimensional proficiency models or tests with multiple attributes. However, the usage of BN in the psychometric field is still in its infancy compared to CDMs and IRT. One important advantage of the BN approach over other measurement approaches is that it can estimate the unknown parameters and hidden structural relationships based on the observed data (Bolstad, 2007). The following section reviews the current studies on the integration of BN modeling in assessment.

Culbertson (2016) reviewed the current state of the BN application in the educational assessment and concluded that BN can diagnostically model content domains. Similarly, Almond et al. (2009) suggested that BN is powerful for providing diagnostic individual-level and group-level information on student proficiency which could help teachers to conduct customized instruction. Researchers have explored ways to combine BN and IRT when analyzing assessment data. For example, Ueno (2002) proposed a BN-IRT model to relax the local independence assumption by explaining the local dependence within the probabilistic network relationship of BN modeling. They found that the proposed BN-IRT model provides better results than the traditional IRT model in terms of model fit. Later in a same condition of detecting local dependence, Hashimoto and Ueno (2011) constructed network relations among items based on BN and computed an index of conditional mutual information to determine the dependence among items.

BN performs well with large data in the context of technology-integrated assessments and an intelligent learning environment. Nouh et al. (2006) presented a computer-based Intelligent Tutoring System (ITS) to diagnose student achievement based on BN and at the same time use IRT to select items adaptive to students of any achievement levels. By using the same ITS tool, Liu (2009) found that even when item responses do not explicitly reflect students' competence, BN can help researchers to identify proficiency model with indirect observations. Specifically, BN can indirectly estimate the competence of students through students' item responses and can learn the structural relationships of attributes. Moreover, the BN method helps teachers and educators to make better decisions on instructional and assessment design.

In a more practical psychometric consideration, researchers have applied BN to the context of computer adaptive testing (CAT) and the large-scale operational assessment. Desmarais and Pu (2005) utilized partial order knowledge structure (POKS), which was also considered as a format of BN model structure, to directly depict the knowledge structure among the attributes measured by the items. They compared the POKS model and the 2-PL IRT model on a CAT exam and concluded that POKS works well with small test data and yields accurate examinee classification in terms of proficiency levels on attributes. In another case, Culbertson and Li (2012) applied BN to analyze the attribute structure for a medical licensing exam. They described the development process of a BN model using the operational data and investigates its measurement precision. Later, Culbertson (2014) conducted a simulation study to examine the performance of BN-based item selection criteria given different conditions of information distribution reflected among attributes for a CAT assessment. He discussed how different BN-based item selection methods impact measurement precision in CAT. Recently, Chen et al. (2018) have proposed a BN-based framework to make recommendations on learning materials for adaptive learning system.

Researchers also used BN to understand the development of the learning process and the relationship among attributes. Pek and Poh (2004) introduced a tutoring system which provides adaptive assistance to students based on a BN modeling of item parameters, students' mastery of the key concepts, and an IRT structure to compute the probability of response correctness. Similarly, Steedle (2008) investigated two contrasting assessment systems: learning progressions and facet-based assessments based on BN modeling.

In the realm of longitudinal assessment data, West et al. (2012) and Choi (2012) constructed a dynamic BN model to investigate the relationships between learning progression and learning tasks. The dynamic BN models, also known as Hidden Markov Model (HMM), are used to model learning progressions over time by making a connection among the associated learning progression theory, assessment design, and inferences on student achievement. The application of dynamic BN is more of interest in formative assessments with a purpose of assessing student learning progress during instruction. They pointed out the advantages of using BN to model learning progressions: (a) it depends on expert input to build the initial conditional dependence among components and then updates the relationships using observed data of real time; (b) it is flexible to model complex data or small sample data; (c) it provides information to help teachers and educators to diagnose students' achievement levels regarding attributes and learning progressions and update their test design.

Lee et al. (2015) applied Bayesian Knowledge Tracing (BKT) algorithm, an algorithm based on BN, to score an interactive learning task. BKT algorithm holds the similar idea with BN model in tracing the learning process based on the dependent relationships among attributes and in inferring whether students master one or multiple components. They can be used to examine students' knowledge acquisition approaches and to provide implications for designing learning tasks. Khajah et al. (2014) integrated BKT algorithm and IRT. Also working with ITS, this study used both MLE and Bayesian approaches to estimate parameters for the BKT, IRT and combined models and concluded that the BKT-IRT model and the IRT model outperform the BKT model when estimating student proficiency. Different from Khajah et al. (2014)'s results, Wilson et al. (2016) discovered that structural IRT is the best-performing model compared to IRT and BKT model and suggested that grouping information is useful to predict student responses while BKT model is more advantageous when considering the inclusion of prior information from experts.

The BN modeling framework corresponds to the principles of ECD. Almond et al. (2007) highlighted the use of BN modeling and its integration with ECD when conducting cognitive diagnosis. They demonstrated the framework of ECD in BN modeling in two steps: the proficiency model, with the conditional dependence among proficiency variables; the evidence model, with the presentation of how observed variables connected with proficiency variables. Further, Wu (2014) found that BN models are slightly better with small samples compared to CDMs as it may rely on the relationships among attributes to estimate parameters in a short test.

Bolt (2007) discussed some advantages and weaknesses of BN modeling. BN is accessible and flexible to construct all aspects of assessment following the ECD framework. However, like CDM, BN modeling relies heavily on expert judgment. Therefore, it may sometimes unclear whether experts can reflect the real relationships between items and attributes, or how real data are used to be representative of their judgment and accurate information.

Taken together, the purpose of applying BN to diagnostic assessment is to classify students into different proficiency levels on each attribute according to their item responses (Rupp, et al., 2008). Under the Bayesian approach, the classification results can be obtained by computing the posterior distribution of parameters after considering the prior information. Further, as Rupp and Templin (2008) stated, BN can be used to specify complex attribute structures and the flexible modeling feature.

2.4 Conceptual Assessment Framework in the Bayesian Network Modeling

BN naturally corresponds to the ECD framework when used to construct an assessment (Almond et al., 2015). ECD intends to gather evidence to make inference on student performance through assessment arguments. Under the ECD framework, the stage of CAF serves as an inferential modeling process transferring the domain conceptualization into a concrete tool to elicit students' knowledge and skills. Rupp et al. (2010) stated that any diagnostic assessment can potentially benefit from the design of the ECD framework.

BN is a graphical modeling procedure providing diagnostic feedback to students' mastery of attributes and explaining structural relationships of cognitive attributes. It can be constructed following the CAF under the ECD framework. Specifically, it demonstrates a path of inference composed of a proficiency model explaining the relationships of proficiency indicators, an evidence model providing statistical inference for the proficiency model from observations, and a task model specifying the features of materials and tasks presented to students. BN can be represented by a graphical structure of variables and a parametric demonstration determining the CPDs for variables in the graphical structure. To echo an assessment framework, the structural part of BN represents the proficiency model, where nodes reflect students' proficiency indicators, and edges are directed arrows specifying the hypothesized or known relationships among indicators. The parametric section allows the mapping from proficiency indicators to observations by parameterizing the relationships and providing inferential evidence to explain the graphical structure based on observed data.

2.5 Summary

In sum, this chapter lays the theoretical and methodological foundations for the study by addressing the topics of assessment frameworks, measurement models, and the connection between them. Specifically, (a) the review of the Assessment Triangle framework and the ECD framework serves as the theoretical foundation for the formative, diagnostic features of the potential BN application results. It specifies the components of the Assessment Triangle, the procedures of ECD, and the three models of CAF (i.e., the proficiency model, the evidence model, the task model). (b) The description of the IRT modeling, CDMs including DINA, DINO and G-DINA, and BN provides an overview of the current popular measurement models. The review of the existing BN application in assessment underlies the methodological support for conducting cognitive diagnosis in this study. Finally, (c) the review of relating BN to CAF demonstrates the natural connection between the theoretical assessment framework and the methodological technique.

Despite the flexibility in modeling and the capacity of providing diagnostic information on student knowledge presented in BN modeling, little research on analyzing assessment data and providing informative diagnosis has considered using BN. It might be attributable to its origin in the artificial intelligence community, which has less connection to the education community. However, research has shown that BN is generally consistent with the assessment framework of ECD and the measurement modeling framework. In addition to these advances, Almond et al. (2015) have listed 10 reasons for considering BN in educational assessment in their book, including its solid mathematical foundations of Bayes' Theorem, its incorporation of theory and experts' option about the measured cognitive domain, its capability of learning observations and optimizing models, its prominent computation time, to name a few. Although limited evidence has presented in the literature on the application of BN in conducting cognitive diagnosis, it has potential to become a prominent conceptual and empirical operationalization of the genuine educational agenda in identifying and realizing students' academic potential. Culbertson (2016)The lack of relevant BN literature and the potential of BN in improving students' learning warrant an empirical comparison of the performance of BN and other measurements models commonly used for cognitive diagnosis including CDMs and a scale-up of the BN application in educational assessment.

2.6 Research Questions

This study aims to answer the following research questions. The first two sets of questions are addressed in a simulation study and the third question is addressed in an analysis of real data:

1. Based on a simulation study, under the different conditions of sample size and test length, how well can the following parameters be recovered: students' mastery profiles, item parameters, person parameters?

2. Based on a stable and optimal selection of sample size and test length (if any) from the first research question,

2a. under the different conditions of attribute types and Q matrix structures, how well can the following parameters be recovered: students' mastery profiles, item parameters, person parameters?

2b. how well the classification accuracies of student attribute mastery can be recovered with a prior information of no structural relationships, partial structural relationships, full structural relationships, and wrong structural relationships? What cognitive diagnostic information can be provided to students and teachers?

2c. how are the models' fitting and accuracies different between the CDM and the BN approach?

3. Based on two existing data sets, how is the effectiveness of the BN approach based on MCMC estimation in analyzing real test data? What cognitive diagnosis can be provided by the BN approach?

3. METHOD

In this chapter, I address the methods used to answer the research questions, including the modeling procedures of BN, MCMC estimation of model parameters, simulation specifications, evaluation criteria, real dataset description, and data analysis plan. First, I explain the conditional distributions defined for proficiency models and evidence models for BN modeling and the joint distribution of all model parameters. Second, I briefly describe the MCMC estimation procedures for this study. Third, I elaborate on the simulation conditions of sample size, test length, Q matrix complexity, attribute types, and prior information on hierarchy in attributes. I also specify the sampling of model parameters for the MCMC estimation. I further explain how the simulation data is generated. Finally, I specify the evaluation criteria of the utility and effectiveness of model results.

3.1 BN Modeling

As demonstrated in Chapter 2, BN models conditional dependence among variables and presents these probabilistic relationships in a graphical display. Its probabilistic computations depend on making a series of Bayesian inferences, which use Bayes’ Theorem to update probabilistic relationships as more information is added. As such, I specify BN modeling in terms of Bayesian inferences and conditional probabilities later in this section. In addition to its flexibility in modeling dependent relationships, the other motivation to promote the application of BN in educational assessment is its natural correspondence with the assessment frameworks of Assessment Triangle (Pellegrino et al., 2001) and ECD (Mislevy & Haertel, 2006). Specifically, the cognition component of BN modeling depends on a proficiency model embodying the theory about the structural relationships among knowledge attributes and the beliefs about students' mastery levels of these attributes. The observation component of BN modeling furnishes the evidence model to elicit students' item responses and make inferences about their mastery status based on the information from the proficiency model. The interpretation component of BN modeling supports the cognitive diagnosis of students' knowledge and strategizes the remediation path for further improvement. Mislevy et al. (1999) have proposed the BN representation of the assessment framework. To be closely aligned with CAF framework in ECD, this study manifests BN modeling in terms of proficiency models and evidence models following their specification.

The proficiency model contains the latent ability variables, which represent students' mastery profile of attributes as a result of the BN application in cognitive diagnosis. They are the target statistical inferences made about examinees. The proficiency model is denoted as $\boldsymbol{\alpha}_{\mathrm{i}} = (\alpha_{\mathrm{i1}}, \dots, \alpha_{\mathrm{iK}})$, indexed for all *N* students (*i* = 1,…,*N*) and all *K* attributes (*k* = 1,…,*K)*. Most of the current literature and studies on cognitive diagnosis focus on classifying students into binary proficiency levels of mastery and non-mastery. However, practically speaking, finer-grained proficiency levels (i.e., more than two levels) could provide teachers and researchers with more information to develop post-diagnosis remediation or learning development plans. To extend the current literature, this study defines students' mastery of attributes on a polytomous scale, which also includes the dichotomous case. Put more generally, α_{i} is a polytomously-valued vector, where $\alpha_{ik} = m$, denoting student *i* reaches the ability level *m*, for *m* = (1,…,*M*), on attribute *k*. For the dichotomous case, *M* is equal to 2.

The inferences made on student mastery of attributes and the structure among *k* attributes are both based on CPDs. To construct the conditional probability for attribute mastery for each student, a prior distribution, which is the distribution of the parameters before any data is observed, is proposed. For the vector of student mastery profile $\boldsymbol{\alpha}_i$ for student *i*, its prior distribution should contain assumptions about the distribution of expected ability levels of each attribute for the class student *i* belongs to in the target population and the structure in attributes through specifying the model structure based on a hyperparameter λ. As such, the probability distribution of the polytomously-valued vector $\boldsymbol{\alpha}$ is expressed as $\theta_{ik} \sim p(\theta|\lambda_k)$, in which $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_K)$. $\boldsymbol{\lambda}_k$ is defined based on the proportion of examinees belonging to each level of the *k*th attribute for level *m* = 1,…,*M*. For example, in a math assessment evaluating three ability levels of each attribute (i.e., *M* = 3), inferences on student ability levels are made in terms of non-mastery, medium mastery, and high mastery. To be congruent with such classification, the Q matrix of this test would contain polytomous values to demonstrate items requiring different ability levels. For example, items may require a medium mastery of multiplication and a high mastery of addition to be answered correctly. Content experts who hold an existing theory or prior experience can provide strong information on the proportions of students distributed across performance levels, which is expressed by $\boldsymbol{\lambda}$, to render a precise prior distribution for $\boldsymbol{\alpha}$. If such information is not available, a vague prior distribution should be provided for $\boldsymbol{\lambda}$ to estimate the probability distribution of $\boldsymbol{\alpha}$. Given that $\boldsymbol{\lambda}$ is unknown in most cases, the distribution of p(λ) can be specified as a Dirichlet distribution, which is a natural conjugate prior distribution for categorically distributed α_{ik}, based on a vague predetermined pseudo count of students distributed across all ability levels for attribute *k*.

Mathematically, for independent attributes with no structural relationships, $\boldsymbol{\alpha}_i = (\alpha_{i1}, \dots, \alpha_{iK})$, $\alpha_{ik} \sim \text{Categorical}(\boldsymbol{\lambda}_k)$, $\boldsymbol{\lambda}_k \sim \text{Dirichlet}\left(\boldsymbol{Count}_{\lambda_k}\right)$, where $\boldsymbol{Count}_{\lambda_k}$ is a vector of

$M_{(k)}$ pseudo counts for attribute *k*. For attributes with dependent relationships, a conditional structure should be incorporated to define α_{ik}. For example, for a test with a dependent relationship as shown in Figure 3, $\alpha_{\mathrm{ik}} \sim \mathrm{Categorical}(\boldsymbol{\lambda}_{\mathrm{k}}), \text{for } k = 1{,}2;$ $\boldsymbol{\lambda}_{\mathrm{k}} \sim \mathrm{Dirichlet}\left(\boldsymbol{Count}_{\lambda_{\mathrm{k}}}\right), \text{for } k = 1{,}2; \theta_{\mathrm{i3}}\left|\left(\alpha_{\mathrm{i1}}, \alpha_{\mathrm{i2}}\right) = c \sim \mathrm{Categorical}(\boldsymbol{\lambda}_{\mathrm{3c}}),\right.$ $\boldsymbol{\lambda}_{\mathrm{3c}} \sim \mathrm{Dirichlet}\left(\boldsymbol{Count}_{\lambda_{\mathrm{3c}}}\right)$, where *c* represents student *i*'s mastery status of attribute 1 and 2, for $c = 1,\ldots,M^2$. In other words, there would be different mastery probabilities of attribute 3 conditional on their mastery status of attributes 1 and 2. As an example, we can set $\lambda_{3,c=3}$ depends on the pseudo counts of students across all ability levels of attribute 3 if they have high mastery of attribute 1 and medium mastery of attribute 2.

The evidence model of BN modeling yields the structure of a probability distribution describing how students' item responses depend on their mastery profile and item parameters. As demonstrated in Chapter 2, a Q matrix predetermined by test developers is used to demonstrate which attributes are measured by each item. In BN modeling, items measuring the same levels of the same attributes are grouped together. Each group of items has its own evidence model indexed by $s = 1,\ldots,S$. Each evidence model elicits item responses $X_{\mathrm{ij(s)}}$ and contributes information about attribute mastery to making inferences for $\boldsymbol{\alpha}$. Student responses are denoted as a matrix $\mathbf{X}_{\mathrm{j}} = \left(X_{\mathrm{j1}}, \ldots, X_{\mathrm{JN}}\right)$ for item $j = 1,\ldots,J$ across all examinees. Regarding item-level parameters, $\boldsymbol{\pi}_{\mathrm{jl}|\boldsymbol{\alpha}\mathrm{i}}$ denotes the conditional probability of responding to item *j* with a value of *l* given students' mastery status $\boldsymbol{\alpha}_{\mathrm{i}}$. That is, $\pi_{\mathrm{jl}|\alpha\mathrm{i}} = P\left(X_{\mathrm{ij}} = x_{\mathrm{ijl}} | \boldsymbol{\alpha}_{\mathrm{i}}\right)$ as described in Levy and Mislevy (2004). More generally, let each mastery status of the attributes measured by each evidence model be labeled by integer $c = 1,\ldots,C$. Then $\boldsymbol{\pi}_{\mathrm{j}|\alpha\mathrm{i}=c_{(\mathrm{s})}}$ represents a vector of conditional probabilities of observing each possible value of X_{ij} on item *j* for students with mastery status *c* on the attributes measured by evidence model *s*.

The distribution of item responses depends on α_{i} and π_{j}, which is denoted as $X_{\mathrm{ij}} \sim p\left(X_{\mathrm{ij}} \middle| \boldsymbol{\alpha}_{\mathrm{i}}, \boldsymbol{\pi}_{\mathrm{j}|\alpha\mathrm{i}=c_{(\mathrm{s})}}\right)$. Further, the distribution of $\boldsymbol{\pi}_{\mathrm{j}|\alpha\mathrm{i}=c_{(\mathrm{s})}}$ has its prior distribution defined by the pseudo count parameter ***Count***$_{\mathrm{sc}}$, which refers to the prior information about the probability of selecting each possible response on the items grouped by evidence model *s* for students with a given mastery status *c*. If ***Count***$_{\mathrm{sc}}$ is unknown, its prior distribution can be determined as *p*(***Count***$_{\mathrm{sc}}$), and ***Count***$_{\mathrm{sc}}$ can be specified as the pseudo counts of students in different mastery status of the associated evidence model *s*. Similar as the Bayesian inferences for $\boldsymbol{\alpha}_{\mathrm{i}}$, $\boldsymbol{\pi}_{\mathrm{j}|\alpha\mathrm{i}=c_{(\mathrm{s})}}$ follows a categorical distribution dependent on ***Count***$_{\mathrm{sc}}$ for polytomous item response or a Bernoulli distribution dependent on ***Count***$_{\mathrm{sc}}$ for dichotomous item responses, which reflects the response behavior on item *j* for students in mastery profile *c*.

Mathematically, $\left(X_{\mathrm{ij}} \middle| \alpha_{\mathrm{ik(s)}} = c\right) \sim \mathrm{Categorical}(\boldsymbol{\pi}_{\mathrm{j}|\alpha\mathrm{i}=c_{(\mathrm{s})}})$ for polytomous responses; $\left(X_{\mathrm{ij}} \middle| \alpha_{\mathrm{ik(s)}} = c\right) \sim \mathrm{Bernoulli}(\boldsymbol{\pi}_{\mathrm{j}|\alpha\mathrm{i}=c_{(\mathrm{s})}})$ for dichotomous responses; $\boldsymbol{\pi}_{\mathrm{j}|\alpha\mathrm{i}=c_{(\mathrm{s})}} \sim \mathrm{Dirichlet}(\boldsymbol{Count}_{\mathrm{sc}}),\ \boldsymbol{\pi}_{\mathrm{j}|\alpha\mathrm{i}=c_{(\mathrm{s})}} = \left(\boldsymbol{\pi}_{\mathrm{j}1}, \ldots, \boldsymbol{\pi}_{\mathrm{jC}}\right)$, where $\alpha_{\mathrm{ik(s)}}$ represents the mastery status of the attribute mastery required by evidence model *s*; $\boldsymbol{Count}_{\mathrm{sc}}$ is a vector with a length *L* reflecting the pseudo counts of each possible value of item *j* for student *i* who belongs to mastery profile *c* of the measured attributes for evidence model *s*.

Put more succinctly, I delineated the π parameters in terms of likelihood functions. I used a dichotomous item as an illustration. Suppose that this item has two attributes as its parents in the DAG, that is, measuring two attributes. For each combination of the parent variables, there is a conditional distribution for the possible responses to this item. Table IV illustrates the conditional probability table as the likelihood function for the item response coded as 0 and 1 that depends on respondents' discrete mastery status on the two latent variables for the measured attributes, each coded as the number of attributes reaching the required mastery levels taking on values of 0, 1, 2. As shown in the table, it can be noted that there are three π parameters associated with item *j*, which depends on two attributes. These π parameters were sampled from Beta distribution based on prior pseudo counts.

TABLE IV CONDITIONAL PROBABILITY TABLE FOR RESPONSES TO ITEM *J* THAT DEPENDS ON TWO ATTRIBUTES

Number of attributes having reached the required mastery levels of the measured attributes	$P(X_j \mid \alpha_1,\alpha_2)$	
	0	1
0	$1-\pi_{j,(00)}$	$\pi_{j,(00)}$
1	$1-\pi_{j,(01,\ 10)}$	$\pi_{j,(01,\ 10)}$
2	$1-\pi_{j,(11)}$	$\pi_{j,(11)}$

Note: $\pi_{j(ab)}$ is the probability of correctness for item *j* When $\alpha_1 = a$ and $\alpha_2 = b$.

After setting up all the modeling pieces, the full joint probability distribution of the responses $X_{(s)ij}$ of *N* students to *J* items nested within *S* evidence models can be expressed as follows according to Bayes' Theorem. It is used to simulate the posterior distributions of the unknown parameters.

$$p(\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\lambda}) = \prod_{s=1}^{S} \prod_{j=1}^{J} \prod_{i=1}^{N} p\left(X_{(s)ij} | \alpha_i, \pi_{(s)j}\right) p\left(\pi_{(s)j}\right) p(\alpha_i | \lambda) p(\lambda) \qquad (12)$$

3.2 MCMC Estimation

Once the response data are observed, the Bayesian inference on each parameter is computed by deriving the posterior distribution from posterior simulation, which is also called Markov Chain Monte Carlo (MCMC). The MCMC procedure follows a Markov chain of estimation iterations, in which the probability of each estimation iteration depends only on the state of parameters estimated in the previous iterations. The MCMC estimation of BN modeling for this study demonstrates that given the observed responses, the posterior distribution is estimated for the unobserved parameters (i.e., $\boldsymbol{\alpha}$, $\boldsymbol{\pi}$, $\boldsymbol{\lambda}$).

In this study, I carried out MCMC based on the approach of Gibbs sampling by using the Bayesian inference Using just another Gibbs sampler (JAGS; Plummer, 2015), and the R2jags package (Su & Yajima, 2015) in R. The R2jags package is used to run the JAGS code on the R platform. The Gibbs sampling procedure for this study is described as follows.

Given $\boldsymbol{Count}_{\lambda_k}, \boldsymbol{Count}_{SC}, \boldsymbol{X}$ are known, and the full joint probability distribution:

$$p(\mathbf{X}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\lambda}) = \prod_{s=1}^{S} \prod_{j=1}^{J} \prod_{i=1}^{N} p\left(X_{(\mathrm{s})\mathrm{ij}} | \alpha_{\mathrm{i}}, \pi_{(\mathrm{s})\mathrm{j}}\right) p\left(\pi_{(\mathrm{s})\mathrm{j}}\right) p(\alpha_{\mathrm{i}} | \lambda) p(\lambda),$$

Step 1. Sample

$\boldsymbol{\lambda}^0 \sim Dirichlet(\boldsymbol{Count}_{\lambda_k})$

$\boldsymbol{\pi}^0 \sim Dirichlet(\ \boldsymbol{Count}_{SC})$

$\boldsymbol{\alpha}^0 \sim Categorical(\boldsymbol{\lambda}^0)$

Step 2. Given observed response data $\boldsymbol{X}$, for t = 0, 1,…,T, where T represents the convergence point, the value of each model parameter is drawn from the following CPDs based on the full probability model:

Draw

$\boldsymbol{\alpha}^{\mathrm{t+1}} \sim p(\boldsymbol{\alpha}|\boldsymbol{X}, \boldsymbol{\pi}^{\mathrm{t}}, \boldsymbol{\lambda}^{\mathrm{t}})$

$= \frac{p(\boldsymbol{\alpha}, \boldsymbol{X}, \boldsymbol{\pi}^{\mathrm{t}}, \boldsymbol{\lambda}^{\mathrm{t}})}{p(\boldsymbol{X}, \boldsymbol{\pi}^{\mathrm{t}}, \boldsymbol{\lambda}^{\mathrm{t}})}$ according to Bayes' Theorem,

$= \frac{p(\boldsymbol{X}|\boldsymbol{\alpha}, \boldsymbol{\pi}^{\mathrm{t}})\, p(\boldsymbol{\pi}^{\mathrm{t}})\, p(\boldsymbol{\alpha}|\boldsymbol{\lambda}^{\mathrm{t}})\, p(\boldsymbol{\lambda}^{\mathrm{t}})}{p(\boldsymbol{X}|\boldsymbol{\pi}^{\mathrm{t}}, \boldsymbol{\lambda}^{\mathrm{t}})\, p(\boldsymbol{\pi}^{\mathrm{t}}|\boldsymbol{\lambda}^{\mathrm{t}})\, p(\boldsymbol{\lambda}^{\mathrm{t}})}$ according to the chain rule on the denominator,

$= \frac{p(\boldsymbol{X}|\boldsymbol{\alpha}, \boldsymbol{\pi}^{\mathrm{t}})\, p(\boldsymbol{\pi}^{\mathrm{t}})\, p(\boldsymbol{\alpha}|\boldsymbol{\lambda}^{\mathrm{t}})\, p(\boldsymbol{\lambda}^{\mathrm{t}})}{p(\boldsymbol{X}|\boldsymbol{\pi}^{\mathrm{t}})\, p(\boldsymbol{\pi}^{\mathrm{t}})\, p(\boldsymbol{\lambda}^{\mathrm{t}})}$ as $\boldsymbol{X}$ is independent of $\boldsymbol{\lambda}$, $\boldsymbol{\pi}$ is independent of $\boldsymbol{\lambda}$,

$= \frac{p(\boldsymbol{X}|\boldsymbol{\alpha}, \boldsymbol{\pi}^{\mathrm{t}})\, p(\boldsymbol{\alpha}|\boldsymbol{\lambda}^{\mathrm{t}})}{p(\boldsymbol{X}|\boldsymbol{\pi}^{\mathrm{t}})}$

Draw

$\boldsymbol{\pi}^{\mathrm{t+1}} \sim p(\boldsymbol{\pi}|\mathbf{X}, \boldsymbol{\alpha}^{\mathrm{t+1}}, \boldsymbol{\lambda}^{\mathrm{t}})$

$= \frac{p(\boldsymbol{\pi}, \boldsymbol{X}, \boldsymbol{\alpha}^{\mathrm{t+1}}, \boldsymbol{\lambda}^{\mathrm{t}})}{p(\boldsymbol{X}, \boldsymbol{\alpha}^{\mathrm{t+1}}, \boldsymbol{\lambda}^{\mathrm{t}})}$ according to Bayes' Theorem,

$= \frac{p(\boldsymbol{X}|\boldsymbol{\alpha}^{\mathrm{t+1}}, \boldsymbol{\pi})\, p(\boldsymbol{\pi})\, p(\boldsymbol{\alpha}^{\mathrm{t+1}}|\boldsymbol{\lambda}^{\mathrm{t}})\, p(\boldsymbol{\lambda}^{\mathrm{t}})}{p(\boldsymbol{X}|\boldsymbol{\alpha}^{\mathrm{t+1}}, \boldsymbol{\lambda}^{\mathrm{t}})\, p(\boldsymbol{\alpha}^{\mathrm{t+1}}|\boldsymbol{\lambda}^{\mathrm{t}})\, p(\boldsymbol{\lambda}^{\mathrm{t}})}$ according to the chain rule on the denominator,

$= \frac{p(\boldsymbol{X}|\boldsymbol{\alpha}^{\mathrm{t+1}}, \boldsymbol{\pi})\, p(\boldsymbol{\pi})\, p(\boldsymbol{\alpha}^{\mathrm{t+1}}|\boldsymbol{\lambda}^{\mathrm{t}})\, p(\boldsymbol{\lambda}^{\mathrm{t}})}{p(\boldsymbol{X}|\boldsymbol{\alpha}^{\mathrm{t+1}})\, p(\boldsymbol{\alpha}^{\mathrm{t+1}}|\boldsymbol{\lambda}^{\mathrm{t}})\, p(\boldsymbol{\lambda}^{\mathrm{t}})}$ as $\boldsymbol{X}$ is independent of $\boldsymbol{\lambda}$,

$= \frac{p(\boldsymbol{X}|\boldsymbol{\alpha}^{\mathrm{t+1}}, \boldsymbol{\pi})\, p(\boldsymbol{\pi})}{p(\boldsymbol{X}|\boldsymbol{\alpha}^{\mathrm{t+1}})}$

Draw

$\boldsymbol{\lambda}^{\mathrm{t+1}} \sim p(\boldsymbol{\lambda}|\mathbf{X}, \boldsymbol{\alpha}^{\mathrm{t+1}}, \boldsymbol{\pi}^{\mathrm{t+1}})$

$$= \frac{p(\boldsymbol{\lambda}, \boldsymbol{X}, \boldsymbol{\alpha}^{\mathrm{t}+1}, \boldsymbol{\pi}^{\mathrm{t}+1})}{p(\boldsymbol{X}, \boldsymbol{\alpha}^{\mathrm{t}+1}, \boldsymbol{\pi}^{\mathrm{t}+1})} \qquad \text{according to Bayes' Theorem,}$$

$$= \frac{p(\boldsymbol{X}|\boldsymbol{\alpha}^{\mathrm{t}+1}, \boldsymbol{\pi}^{\mathrm{t}+1})\, p(\boldsymbol{\pi}^{\mathrm{t}+1})\, p(\boldsymbol{\alpha}^{\mathrm{t}+1}|\boldsymbol{\lambda})\, p(\boldsymbol{\lambda})}{p(\boldsymbol{X}|\boldsymbol{\alpha}^{\mathrm{t}+1}, \boldsymbol{\pi}^{\mathrm{t}+1})\, p(\boldsymbol{\alpha}^{\mathrm{t}+1}|\boldsymbol{\pi}^{\mathrm{t}+1})\, p(\boldsymbol{\pi}^{\mathrm{t}+1})} \qquad \text{according to the chain rule on the denominator,}$$

$$= \frac{p(\boldsymbol{X}|\boldsymbol{\alpha}^{\mathrm{t}+1}, \boldsymbol{\pi}^{\mathrm{t}+1})\, p(\boldsymbol{\pi}^{\mathrm{t}+1})\, p(\boldsymbol{\alpha}^{\mathrm{t}+1}|\boldsymbol{\lambda})\, p(\boldsymbol{\lambda})}{p(\boldsymbol{X}|\boldsymbol{\alpha}^{\mathrm{t}+1}, \boldsymbol{\pi}^{\mathrm{t}+1})\, p(\boldsymbol{\alpha}^{\mathrm{t}+1})\, p(\boldsymbol{\pi}^{\mathrm{t}+1})} \qquad as\ \boldsymbol{\alpha} \text{ is independent of } \boldsymbol{\pi},$$

$$= \frac{p(\boldsymbol{\alpha}^{\mathrm{t}+1}|\boldsymbol{\lambda})\, p(\boldsymbol{\lambda})}{p(\boldsymbol{\alpha}^{\mathrm{t}+1})}$$

The Empirical Bayes method was used to estimate hyperparameters in the above CPDs and to further approximate the distribution for each model parameter. As the starting values of model parameters are randomly sampled or chosen for convenience, they are usually far from the target values. The posterior estimation of model parameters would approximate the target values when the number of iterations increase. When the Markov chain reaches its convergence and the parameter estimation becomes stable, the posterior distribution at this iteration demonstrates the target model parameter estimation. To make the Markov chain reaches the target values more quickly, I dropped some iterations at the beginning of MCMC as burn-ins.

3.3 Simulation Specification

3.3.1 Simulation Factors

As the application of BN in cognitive diagnosis is still new to the field, it is necessary to evaluate how different assessment scenarios or contexts impact its estimation. I varied the following factors in this simulation study: (a) sample size, (b) test length, (c) Q matrix complexity, (d) attribute type, and (e) the structure in attributes. These factors were chosen according to the current challenges or contexts of conducting cognitive diagnosis. First, cognitive diagnosis is usually noted to be sensitive to sample size because it is hard to make reliable estimation of model parameters for small sample sizes (Chiu et al., 2018). For this reason, most studies were conducted for sample sizes larger than 500. However, for some small education programs or interventions where a large sample size is not feasible to achieve, it may hinder the applicability of cognitive diagnosis in some small-scale classroom contexts. It is therefore necessary to examine whether the BN approach, as an alternative way of cognitive diagnosis, could alleviate this constraint and provide informative feedback regarding student proficiency in broader educational contexts. To achieve this goal, I examined sample sizes of 50, 100, 500. Another factor that accounts for data size is the test length. Almond et al. (2015) claimed that, in addition to a larger sample size, increasing the test length would also help to produce more accurate estimates for BN models. However, it is not clear how sample size, test length, and its interaction impact parameter estimation results of BN modeling. It is therefore necessary to probe whether test length may influence the effectiveness of the BN approach. For this purpose, I examined the test lengths of 15 and 30 items in this study.

I also evaluated the performance of the BN approach under different structures of the Q matrix. As demonstrated in Chapter 2, tests can be designed based on a simple structure or a complex structure given the requirements of test blueprint and test purposes. In a simple structure Q matrix, only one attribute is measured by each item, while more than one attribute can be loaded on one item in a complex structure Q matrix. Additionally, in current literature, most CDM models were developed to provide binary classification on student mastery of attributes and fail to accommodate polytomous attributes (Chen & de la Torre, 2013). However, finer-grained inferences regarding student mastery profile would help instructors and researchers to customize their instruction and intervention according to different learning demands. This study investigates the application of the BN approach in classifying students into more mastery levels based on polytomous attributes, which may extend the utility of current cognitive diagnosis. Finally, as one objective of this study is to investigate how BN performs for assessment with a structure in attributes and how it recovers the structural relationships, I also varied the prior information on the structure in attributes by inputting no information, partial information, and full information. The simulation scenarios are summarized in Table V.

TABLE V SIMULATION FACTORS AND LEVELS

Factor	# of conditions	Level
Sample size	3	50, 100, 500
Test length	2	15, 30
Q matrix complexity	2	Simple, Complex
Attribute type	2	Dichotomous, Polytomous
Prior info on attribute structure	4	No info, Partial info, Full info, Wrong info

3.3.2 Q Matrix

Q matrix is used to demonstrate the measured attributes by each item. In order to design the simulation of this study with theoretical rationale rather than completely from scratch, I have extended the Q matrix from the fraction-subtraction test of Tatsuoka’s (1983) study based on its item distributions and revised the structural structure proposed by Sinharay et al. (2004). That said, the tests simulated for this study has five attributes containing a structural relationship.

Figure 4 depicts the structural relationships among attributes used in this study after adaption. Tables VI-IX describe the Q matrices used in this study under different scenarios. Table VI and Table VII are simple structure Q matrices, where each item only measures one attribute, and the attributes in this simple structure test were set to be evenly distributed. Table VIII and Table IX are complex structure Q matrices, where each item may measure more than one attribute, and the distribution of the items on each attribute follows that of the original Q matrix for the Tatsuoka (1984) mixed number subtraction test. Additionally, Table VIII and Table IX contain polytomous attributes, in which I balanced the distribution of the three levels based on the hierarchy in attributes. The mastery requirement for an item should be associated with the medium or high mastery of its prerequisite attributes. For example, item 9 requires high mastery of attributes 1 and 5 but medium mastery of attributes 3 and 4 as attributes 1 and 5 are prerequisites for attributes 3 and 4.

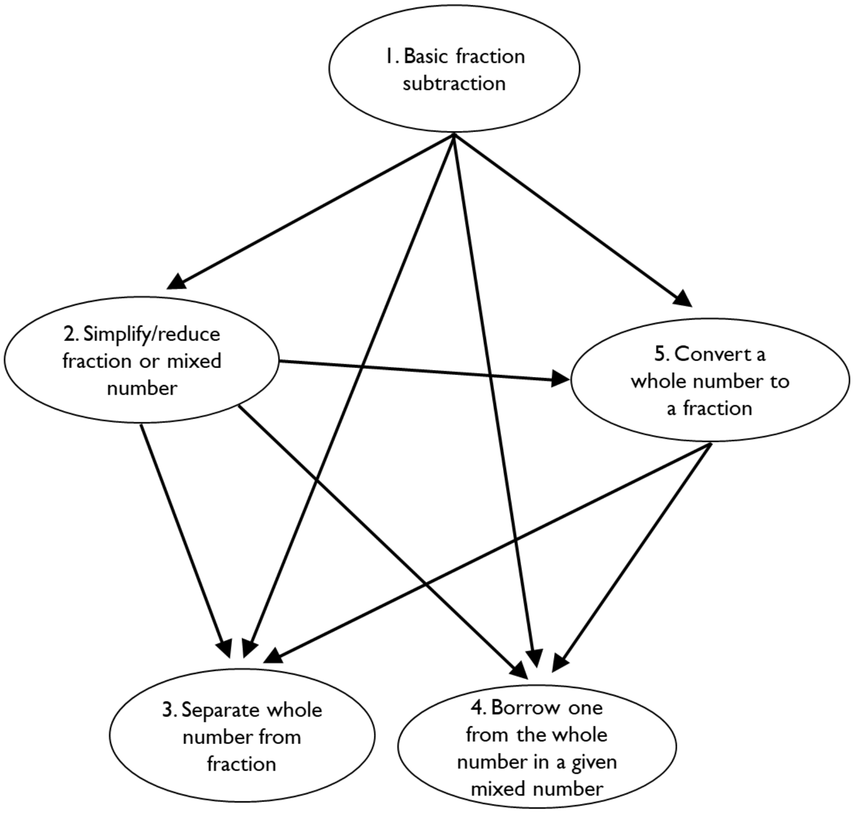

1. Basic fraction subtraction
2. Simplify/reduce fraction or mixed number
5. Convert a whole number to a fraction
3. Separate whole number from fraction
4. Borrow one from the whole number in a given mixed number

Figure 4. The structural relationships among the attributes used in the simulation. This structure was adapted from the structure from Sinharay et al. (2004).

TABLE VI SIMPLE STRUCTURE Q MATRIX WITH DICHOTOMOUS ATTRIBUTES

Evidence Model	Item No. of different test lengths		Attributes				
	15	30	1	2	3	4	5
1	1-2	1-4	1				
2	3-4	5-8		1			
3	5-7	9-14			1		
4	8-10	15-20				1	
5	11-15	21-30					1

Note. The measured attributes measured by each item are labeled as 1.

TABLE VII COMPLEX STRUCTURE Q MATRIX WITH DICHOTOMOUS ATTRIBUTES

Evidence Model	Item No. of different test lengths		Attributes				
	15	30	1	2	3	4	5
1	1-2	1-4	1				
2	3	5-6	1	1			
3	4-6	7-12	1		1		
4	7-11	13-22	1		1	1	
5	12-14	23-28	1		1	1	1
6	15	29-30	1	1	1	1	

Note. The measured attributes measured by each item are labeled as 1.

TABLE VIII SIMPLE STRUCTURE Q MATRIX WITH POLYTOMOUS ATTRIBUTES

Evidence Model	Item No. of different test lengths		Attributes				
	15	30	1	2	3	4	5
1	1	1-2	1				
2	2	3-4	2				
3	3	5-6		1			
4	4	7-8		2			
5	5-6	9-12			1		
6	7	12-14			2		
7	8	15-16				1	
8	9-10	17-20				2	
9	11-13	21-26					1
10	14-15	27-30					2

Note. The measured attributes requiring high mastery level are labeled as 2, medium mastery are labeled as 1.

TABLE IX COMPLEX STRUCTURE Q MATRIX WITH POLYTOMOUS ATTRIBUTES

Evidence Model	Item No. of different test lengths		Attributes				
	15	30	1	2	3	4	5
1	1	1-2	1				
2	2	3-4	2				
3	3	5-6	2	1			
4	4-5	7-10	2		1		
5	6	11-12	2		2		
6	7-8	13-16	2		1	1	
7	9-10	17-20	2		2	1	
8	11	21-22	2		2	2	
9	12	23-24	2		1	1	2
10	13	25-26	1		1	1	1
11	14	27-28	2		1	1	1
12	15	29-30	2	2	1	1	

Note. The measured attributes requiring high mastery level are labeled as 2, medium mastery are labeled as 1.

3.3.3 Parameter Sampling

This section specifies the prior distributions for person and item parameters to sample from in the simulation based on the specified structural relationships as shown in Figure 4.

Generally speaking, the posterior distribution of parameters in Bayesian estimation relies on prior distribution and the observed data. Stronger prior information would result in posterior distribution estimation closer to prior distribution. Therefore, if reliable information on prior distribution is not available, no or weak prior distribution should be provided so that it has little influence on the poster distribution of model parameters. Considering that this study examines the application of BN modeling in a relatively small sample size, the estimation accuracy might be influenced if no information was provided (i.e., same priors for each parameter). For this reason, when conducting the MCMC estimation for model parameters using JAGS in this study, I used weak priors, which reflects a lack of accurate prior knowledge but with some vague information about the proportions of student classes of different mastery profiles.

The sampling of person parameters is guided by the proficiency model. According to the dependent relationships among the five attributes, the proficiency model can be factorized as:

$$p(\boldsymbol{\alpha}_{\mathrm{i}}) = p(\alpha_{\mathrm{i}3}|\alpha_{\mathrm{i}1}, \alpha_{\mathrm{i}2}, \alpha_{\mathrm{i}5})p(\alpha_{\mathrm{i}4}|\alpha_{\mathrm{i}1}, \alpha_{\mathrm{i}2}, \alpha_{\mathrm{i}5})p(\alpha_{\mathrm{i}5}|\alpha_{\mathrm{i}1}, \alpha_{\mathrm{i}2})p(\alpha_{\mathrm{i}2}|\alpha_{\mathrm{i}1})p(\alpha_{\mathrm{i}1}) \quad (13)$$

Specifically, this proficiency model demonstrates that attribute 1 is the prerequisite for all the other attributes. Attribute 2 is only dependent on attribute 1. To learn attribute 5, students usually first master both attributes 1 and 2. Attributes 1, 2, 5 are prerequisites to attributes 3 and 4 respectively. Following this relationship, **λ** can be defined as follows:

$\lambda_1 = p(\alpha_1 = m_1)$ for $m_1 = 0, \dots, M$

$\lambda_2 = p(\alpha_2 = m_2|\alpha_1 = m_1)$ for $m_2 = 0, \dots, M$

$\lambda_5 = p(\alpha_5 = m_5|\alpha_2 = m_2, \alpha_1 = m_1)$ for $m_5 = 0, \dots, M$

$\lambda_3 = p(\alpha_3 = m_3|\alpha_5 = m_5, \alpha_2 = m_2, \alpha_1 = m_1)$ for $m_3 = 0, \dots, M$

$\lambda_4 = p(\alpha_4 = m_4|\alpha_5 = m_5, \alpha_2 = m_2, \alpha_1 = m_1)$ for $m_4 = 0, \dots, M.$

When selecting the starting values of the pseudo counts for the conjugate prior distribution for the distribution of mastery levels for each attribute, the larger sum of the pseudo counts for each mastery level would provide stronger prior knowledge about **λ**, while a smaller sum would have less influence on the subsequent posterior simulation (i.e., MCMC estimation). As there is no prior knowledge about the distribution of mastery profiles for this study, I used a small sum of pseudo counts, 10, and a seemingly reasonable distribution for the mastery levels of each attribute to maintain a weak prior information and avoid much influence on posterior estimation.

For an assessment with two ability levels, the conjugate prior distribution for mastery levels of each attribute follows a beta distribution. According to the structural relationships, attribute 1 tends to be a basic attribute for the target population to master. Therefore, I assumed 80% ($= \frac{8}{8+2}$) of students are expected to master attribute 1, and λ_1 follow Beta (8,2). For the rest attributes which are all dependent on other attributes, there would be multiple $\boldsymbol{\lambda}$ associated with each attribute specifying the probability of mastery based on the mastery status of prerequisite attributes. For example, attribute 5 has attributes 1 and 2 as prerequisites and its probabilities of mastery can be categorized into groups of students who master none of the two attributes, who master one of the two attributes, and those who master both attributes. In this case, there are three λ parameters associated with attribute 5. All the pseudo counts for sampling the λ parameters for dichotomous attributes are summarized in Table X.

TABLE X PRIOR DISTRIBUTIONS FOR PERSON PARAMETERS (Λ) FOR DICHOTOMOUS ATTRIBUTES IN THE SIMULATION

	Mastery Status of Prerequisite Attributes			
	None	One	Two	Three
Attribute 1		Beta (8,2)		
Attribute 2	Beta (2,8)	Beta (8,2)		
Attribute 3	Beta (2,8)	Beta (4,6)	Beta (6,4)	Beta (8,2)
Attribute 4	Beta (2,8)	Beta (4,6)	Beta (6,4)	Beta (8,2)
Attribute 5	Beta (2,8)	Beta (5,5)	Beta (8,2)	

For an assessment with three ability levels, $\boldsymbol{\lambda}$ follow Dirichlet distributions. We can select the pseudo count parameter with a sum of 10 and distributed as (5, 3, 2) as a standard. It suggests that, given their mastery of parent attributes to the current attribute, 50% ($= \frac{5}{5+3+2}$) of students in the target population tend to fully acquire the current measured attribute, 30% ($= \frac{3}{5+3+2}$) partially acquire the measured domain, and 20% ($= \frac{2}{5+3+2}$) fail to acquire the measured attribute. For example, following the structural relationships among the attributes in the simulated assessments, $\boldsymbol{\lambda}_1 \sim \mathrm{Dirichlet}(5, 3, 2)$ reflects that 50% in the population master the first attribute in the high level, 30% in the medium, 20% in the non-mastery. $\boldsymbol{\lambda}_{2,0} \sim \mathrm{Dirichlet}(2, 3, 5)$ reflects that, given that students failed to master the first attribute, 20% in the population would master the second attribute in the high level, 30% in the medium, 50% non-mastery. The subscript notation of λ before the comma represents the attribute for this λ, and the notation after the comma represents the mastery level of this attribute's parent attributes. For a more complex case, I may need to adjust the standard pseudo count parameters to match the need. For example, $\boldsymbol{\lambda}_{5,21} \sim \mathrm{Dirichlet}(6, 2, 2)$ means that if students have mastered attribute 1 in a high level and attribute 2 in a medium level, they tend to have 60% probability of mastering attribute 5 in the high level, 20% probability in the medium and non-mastery. In this case, I increased the probability of correctness, which is represented by the first pseudo count, to match this mastery profile. Similar procedures will be conducted for attribute 3 and 4. All the pseudo counts for sampling the λ parameters from Dirichlet distribution are summarized in Table XI.

TABLE XI PRIOR DISTRIBUTIONS FOR PERSON PARAMETERS (Λ) FOR POLYTOMOUS ATTRIBUTES IN SIMULATION

	Sum of mastery levels on prerequisite attributes						
	None	One	Two	Three	Four	Five	Six
Attribute 1		Dirichlet (1,2,7)					
Attribute 2	Dirichlet (7,2,1)	Dirichlet (3,5,2)	Dirichlet (1,2,7)				
Attribute 3	Dirichlet (7,2,1)	Dirichlet (6,2,2)	Dirichlet (5,3,2)	Dirichlet (3,5,2)	Dirichlet (1,5,4)	Dirichlet (1,3,6)	Dirichlet (1,2,7)
Attribute 4	Dirichlet (7,2,1)	Dirichlet (6,2,2)	Dirichlet (5,3,2)	Dirichlet (3,5,2)	Dirichlet (1,5,4)	Dirichlet (1,3,6)	Dirichlet (1,2,7)
Attribute 5	Dirichlet (7,2,1)	Dirichlet (2,6,2)	Dirichlet (3,5,2)	Dirichlet (1,5,4)	Dirichlet (1,2,7)		

Based on the prior distribution of $\boldsymbol{\lambda}$ from either Beta distribution or Dirichlet distribution, the distribution of $\boldsymbol{\alpha}$ can be specified as

$$\boldsymbol{\alpha}_{\mathbf{1}} \sim \text{Categorical}(\boldsymbol{\lambda}_{\mathbf{1}}),$$

$$\boldsymbol{\alpha}_{\mathbf{2}} \sim \text{Categorical}(\boldsymbol{\lambda}_{\mathbf{2}}),$$

$$\boldsymbol{\alpha}_{\mathbf{3}} \sim \text{Categorical}(\boldsymbol{\lambda}_{\mathbf{3}}),$$

$$\boldsymbol{\alpha}_{\mathbf{4}} \sim \text{Categorical}(\boldsymbol{\lambda}_{\mathbf{4}}),$$

$$\boldsymbol{\alpha}_{\mathbf{5}} \sim \text{Categorical}(\boldsymbol{\lambda}_{\mathbf{5}}).$$

Guided by evidence models, item parameters $\boldsymbol{\pi}$ are defined as the distribution of each possible value an item can take given each mastery profile. Each evidence model has ***Count***$_{\mathrm{sc}}$ to define the conjugate prior distribution for the conditional probabilities $\boldsymbol{\pi}_{\mathrm{j}|\alpha\mathrm{i}=c_{(s)}}$. For example, let the exam has dichotomous items (i.e., L = 2). For the test with complex polytomous attributes, the ***Count***$_{(7)9\mathrm{c}}$ denoting the pseudo counts of possible values (i.e., 0 or 1) for item 9 in evidence model 7 can be specified as (2, 8) for examinees who have mastered the high level of attribute 1 but have not mastered attribute 3 and 4. Note that $\boldsymbol{Count}_{sc}$ are specified as the same across all items grouped within the same evidence model. Based on the $\boldsymbol{Count}_{(7)9c}$, the item parameter $\pi_{9|c}$ for examinees of mastery profile *c* follows Beta ($\boldsymbol{Count}_{(7)9c}$). Further, the item response of examinees in this mastery profile for item 9 would follow Bernoulli ($\boldsymbol{\pi}_{9|c}$). The pseudo counts for prior distribution of $\boldsymbol{\pi}$ parameters for students of each mastery profile are summarized in Table XII.

TABLE XII PRIOR DISTRIBUTIONS FOR DICHOTOMOUS ITEM PARAMETERS (Π) IN THE SIMULATION

Evidence Models (number of measured attributes)	Number of attributes having reached the required mastery levels				
	0	1	2	3	4
1	Beta (2,8)	Beta (8,2)			
2	Beta (2,8)	Beta (5,5)	Beta (8,2)		
3	Beta (2,8)	Beta (4,6)	Beta (6,4)	Beta (8,2)	
4	Beta (2,8)	Beta (3.5,6.5)	Beta (5,5)	Beta (6.5,3.5)	Beta (8,2)

3.3.4 Data Generation and Real Data Description

I took two steps to generate simulated data: the first step is to generate students' mastery profiles depending on the proficiency model, and the second step is to generate dichotomous item responses depending on the item parameters from evidence models. For the first step of generating proficiency model variables, the probabilities ($\boldsymbol{p}$) of the three mastery levels of each attribute among the target population was set to sample from the conjugate prior distribution based on pseudo counts with a sum of 1000. By doing so, the generated parameter would be distributed concentratedly around a specific value with small variance. In this way, replicating each simulation based on similar parameter values would not vary a lot, which is helpful in testing the stability of BN modeling. Based on these probabilities, the mastery levels of each attribute for each student are categorically distributed, that is $\alpha_{\mathrm{ik}} \sim \mathrm{Categorical}(\boldsymbol{p})$. For the second step, the conditional probabilities $\boldsymbol{\pi}_{\mathrm{j}}$ as item parameters are dependent on mastery profiles and the attributes required by the evidence models. Specifically, if a mastery profile *c* has reached the required proficiency levels of the attributes specified by the evidence model *s*, $\boldsymbol{\pi}_{\mathrm{j}}$ for this mastery profile *c* are higher than other mastery profiles. Similarly, the pseudo counts to sample the π parameters have a sum of 1000 so that it would yield stable parameter values across replications for each simulation. Further, depending on the conditional probabilities $\boldsymbol{\pi}_{\mathrm{j}}$, the dichotomous item responses X_{ij} for item *j* for student *i* of mastery profile *c* follows Bernoulli ($\boldsymbol{\pi}_{\mathrm{jc}}$).

The real data used in this study comes from Trends in International Mathematics and Science Study (TIMSS) 2003 Grade 8 Mathematics test carried out by the International Association for the Evaluation of Educational Achievement. It includes 757 examinees and their responses to 21 dichotomously scored items, including 19 multiple-choice questions and two constructed response questions.

3.3.5 Evaluation Criteria

The model performance through simulations is usually evaluated by making comparisons of the estimated results with true known parameters or through comparisons with results from other measurement models based on a set of evaluation criteria. In this section, I explain the evaluation criteria for the effectiveness and utility of BN model results for this study.

Overall model fit. The model fit indices include deviance and posterior predictive modeling checking (PPMC). The deviance is a model fit statistic representing the amount of unexplained variance in the applied model – the higher the value the less accurate the model. It sums the squares of residuals between the predicted outcome and the actual outcome and provide a measure of the total residuals for each model.

The PPMC (Gelman et al., 2013) is used to evaluate the absolute model–data fit. The approach is to replicate observed data or predict data based on the probability function stated in Equation (10). If the function is the correct probability function, then the replicated item responses should be similar to the observed responses. In this study, the discrepancy of the sum of the squared Pearson residuals for person *i* and item *j* (Yan, Mislevy, & Almond, 2003) between the predicted and the observed responses is calculated to evaluate the overall fit of BN modeling. The formula for the sum of the squared Pearson residuals is specified as follows:

$$S\left(Y_{ij};\alpha_i\right) = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(\frac{Y_{ij}-p_{ij}}{\sqrt{p_{ij}(1-p_{ij})}}\right)^2, \tag{14}$$

Where p_{ij} is defined as the probability of answering item *j* correctly for student *i*,

Y_{ij} represents the response to item *j* for student *i.*

The discrepancy measure then represents the posterior predictive probability (PPP) values. When the PPP values are around 0.5, it indicates no systematic distinctions between observed item responses and the item responses predicted by the applied model. In this case, this model has an adequate fit to the data. In contrast, when PPP values are close to 0 or 1 (i.e., PPP value < 0.05 or PPP value > 0.95), it suggests a model misfit.

Then the PPP value of the overall fit represents the proportion of time the observed data is larger than the replicated data. Based on the same idea, the PPP values for item fit and person fit are also calculated. Specifically, item fit is calculated by the *RMSE* of the squared Pearson residuals *S* for each item, and person fit is calculated by the *RMSE* of *S* for each person. In this case, the posterior mean of both *RMSE*s represent the magnitude of discrepancy between observed and predicted responses. Then, the PPP values for each item and each person is calculated by the proportion of time observed data is larger than replicated data.

When comparing the CDM model with the BN approach, deviance was used to indicate the model fitting performance. Across all model fit indices, the CDM model usually gives out AIC, BIC, deviance, while the BN approach, as a Bayesian estimation approach, often provides DIC and deviance. In this case, we can only use deviance as the model fitting criteria for the comparison.

The Evidence model results: item parameters. The item parameters π in BN modeling are interpreted as the probability of solving an item correctly given students' mastery profile on the measured attributes. The procedure of obtaining estimated π was elaborated in Section 3.2. The performance of BN modeling in terms of the estimated item parameters can be evaluated by accuracy, which compares the true item parameters with the estimated item parameters. Researchers (e.g., Choi, 2012; Culbertson, 2014) have reported indices of accuracy: the mean of the bias of the estimated item parameters relative to the true item parameters and the root mean squared standard deviations (*RMSD*s). Biases and *RMSD*s are computed based on Equations 15 and 16.

$$Bias(\hat{\pi}) = \frac{\sum_{r=1}^{N}(\hat{\pi}_r - \pi)}{N}, \quad (15)$$

$$RMSD(\hat{\pi}) = \sqrt{\frac{1}{N}\sum_{r=1}^{N}(\hat{\pi}_r - \pi)^2}, \quad (16)$$

where r denotes the r-th replication, N denotes the number of replications;

$\hat{\pi}$ denotes the estimated item parameter, π denotes the true item parameter.

The Proficiency model results: person parameters and mastery profile. The person parameter λ of BN modeling refers to the probability of a student mastering each attribute. The procedure of obtaining estimated λ was elaborated in Section 3.2. Similar to the item parameters, bias and *RMSD*s are considered as the evaluation criteria to examine ability estimation accuracy as shown in Equations (17), (18), respectively.

$$Bias(\hat{\lambda}) = \frac{\sum_{r=1}^{N}(\hat{\lambda}_r - \lambda)}{N}, \quad (17)$$

$$RMSD(\hat{\lambda}) = \sqrt{\frac{1}{\mathrm{N}}\sum_{\mathrm{r}=1}^{\mathrm{N}}(\hat{\lambda}_{\mathrm{r}} - \lambda)^2}, \quad (18)$$

where r denotes the r-th replication, N denotes the number of replications;

$\hat{\lambda}$ denotes the estimated probability of reaching a mastery level for an attribute,

λ denotes the true probability of reaching a mastery level for an attribute.

In addition to person parameters, the typical results of BN modeling are person classification results. Researcher (e.g., Levy & Mislevy, 2004; Liu & Cheng, 2018; Shu, Henson,

& Willse, 2013) have investigated the classification accuracy between true and estimated results and classification consistency of results across several simulation scenarios. The weighted Cohen's Kappa (see in Equation 19) for the discrete skills (Cohen, 1960) is used to evaluate classification accuracy and consistency of person mastery profile. According to Landis & Koch (1977), the rule of thumb for Cohen's Kappa statistic is that values smaller or equal to 0 suggest no agreement, values between 0.01 and 0.20 suggest none to small agreement, values between 0.21 and 0.40 suggest fair agreement, values between 0.41 and 0.60 suggest moderate agreement, values between 0.61 and 0.80 suggest substantial agreement, and values between 0.81 and 1.00 suggest almost perfect agreement.

$$\kappa = \frac{p_{\mathrm{o(w)}} - p_{\mathrm{e(w)}}}{1 - p_{\mathrm{e(w)}}} \tag{19}$$

where

p_{o} represents the relative observed agreement among the two estimates after being weighted,

p_e represents the hypothetical probability of chance agreement after being weighted.

4. RESULTS

This section presents the results of each research question. First in Section 4.1, I unpack the data generation process and the simulated data analysis results for the first four research questions. Second in Section 4.2, I discuss the results of real data analysis of two existing datasets.

4.1 Simulation Data and Analysis Results

4.1.1 MCMC Convergence

Equation (20) gives the full joint probability distribution over all data and parameters. Once item responses **X** are observed, Bayesian inferences regarding parameters can be obtained by estimating the posterior estimation. The posterior distribution for $\boldsymbol{\alpha}$, π, λ is given as follows:

$$p(\boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\lambda}|\mathbf{X}) = \prod_{s=1}^{S} \prod_{j=1}^{J} \prod_{i=1}^{N} p\left(X_{(\mathrm{s})\mathrm{ij}}|\alpha_{\mathrm{i}}, \pi_{(\mathrm{s})\mathrm{j}}\right) p\left(\pi_{(\mathrm{s})\mathrm{j}}\right) p(\alpha_{\mathrm{i}}|\lambda) p(\lambda) \qquad (20)$$

I first conducted a convergence check to estimate an adequate number of iterations to achieve convergence for the simulation study. To answer the research questions of this study, I checked item and person parameters for convergence. In the convergence check, I chose two Markov chains with 2000 iterations per chain, and used the first 25% of iterations in each chain as burn-ins. I then set the thinning interval to 1, indicating no thinning. Through the MCMC estimation, the JAGS would compute the potential scale reduction factor (Brooks & Gelman, 1998), $\hat{R}$, to assess the convergence of each parameter. According to (Brooks & Gelman, 1998), the $\hat{R}$ value associated with each parameter less than 1.2 indicates its convergence. The preliminary convergence check results have showed that at least 2,000 iterations are necessary for convergence to ensure $\hat{R} < 1.2$ for all item and person parameters. Also, as the preliminary check has achieved convergence, the choice of burn-in value is acceptable (Meyn & Tweedie, 2012). Accordingly, I used the same setup of Bayesian analysis (two Markov chains, one thinning interval, 25% of burn-in) for all the following simulation analyses and results. Although 2000 iterations have already reached convergence, I used 5000 iterations for simulation analysis from a conservative perspective.

For each simulation condition, I produced 20 datasets (i.e., replications) to avoid extreme values occurred in few replications. In most real-world situations, teachers and researchers would at least have a basic idea of the mastery difficulty of each attribute. As such, I used weak priors for person and item parameters instead of priors with no information. By doing so, it aligns with the situation where teachers know some but not all accurate information about the measured attributes. The specific prior assignment can be found in Section 3.3.4.

4.1.2 Research Question 1

Based on a simulation study, under the different conditions of sample size and test length, how well can the following parameters be recovered: students' mastery profiles, item parameters, person parameters?

This research question is aimed to dismantle how sample size and test length may affect parameter estimation when making Bayesian inferences for diagnostic assessments. To make a thorough comparison across different conditions, I used the combination of three sample sizes (*N* = 50, 100, 500) and two test lengths (*J* = 15, 30) on different conditions of Q matrix complexity (i.e., simple and complex) and attribute types (dichotomous and polytomous). To emphasize the impact of sample size and test length, BN modeling results were presented for each condition: the simple dichotomous case, the complex dichotomous case, the simple polytomous case, and the complex polytomous case. The results of each condition across sample sizes and test lengths were evaluated in terms of the model fit index of PPMC which checks how BN modeling fits the data, the biases and *RMSD*s of item and person parameters estimated by BN modeling, the classification accuracy of person attribute mastery measured by the weighted Kappa index. The average PPMC index across all replications was used to demonstrate the model fit for each condition of each case. Consistently, the average bias and *RMSD* of each parameter across all replications for each condition of each case was presented. The weighted Kappa index for each condition of each case was calculated by taking the average of all Kappa values across all replications.

Simple dichotomous case. As shown in Table XIII, all the PPMC indexes meet the fit criteria, , as they all larger than .05 or smaller than .95. They suggest that the BN model fits the simple dichotomous data adequately, therefore the estimated parameters are reliable for analysis. Each item parameter represents the probability of correctness of each item for students with a certain mastery status of the measured attributes and produces generally small biases when compared to the true item parameters that are used to generate simulation data. A decreasing trend of the magnitude of biases is found from the small sample size with the short test length (= .047) to the large sample size with the long test length (= .021). It suggests that if we distribute a longer test length to more students, BN model would yield item parameters closer to true parameters. However, the difference between the shortest test length with the smallest sample size and the longest test length with the largest sample size is around 0.02. This difference can be regarded as trivial if we consider that the assessment scale of 50 examinees and 15 items is low-stakes in a classroom context. In this case, there is neither grossly large biases associated with small sample size or short test length nor apparently small biases associated with large sample size or long test length for item parameter estimation when using BN modeling. The values of *RMSD* associated with each bias also demonstrate a decreasing trend from .010 to .002 with a very small distinction (<= .01) among conditions.

TABLE XIII THE SIMULATION RESULTS FOR THE SIMPLE DICHOTOMOUS CASE

Sample Size	Test Length	Model Fit PPMC	Estimated vs True Item Parameters π		Estimated vs True Person Parameters λ		Classification Accuracy κ
			Absolute-valued Bias	*RMSD*	Absolute -valued Bias	*RMSD*	
50	15	0.232	0.047	0.010	0.052	0.017	0.740
	30	0.101	0.046	0.009	0.048	0.014	0.889
100	15	0.214	0.041	0.007	0.048	0.014	0.784
	30	0.089	0.040	0.006	0.047	0.011	0.901
500	15	0.287	0.026	0.003	0.044	0.009	0.783
	30	0.190	0.021	0.002	0.035	0.005	0.903

The person parameters demonstrate the proportions of students who may master a certain attribute given their mastery status of the prerequisite attributes, which is explained by the structural relationships among attributes in Figure 4. Consistent with the item parameters, we found a decreasing tendency of biases and the associated *RMSD*s across conditions of sample sizes and test lengths. The person parameters tend to be closer to true person parameters in an assessment with a larger sample size and a longer test length. The largest difference in the biases across conditions, which is also considered as small, is between the condition of 50 students with 15 items (= .052) and 500 students with 30 items (=.035). The range of the variety in the associated *RMSD*s is also trivial (Δ =.012). These results highlight the little impact of sample size and test length on person parameters estimated by BN modeling in the simple dichotomous case. The weighted Kappa index is calculated to evaluate the agreement between the estimated mastery classification and the true classification on each attribute for each student. As a result, the classification accuracies for the conditions with 30 items outperform the conditions with 15 items regardless of sample sizes as shown in Figure 5. According to the rule of thumb for the Kappa statistic, the classification accuracies for the conditions of 30 items have reached perfect agreement and the accuracies for 15 items have had substantial agreement. In general, the BN approach performs well on the classification accuracy for attribute mastery, with longer test length showing slightly higher agreement.

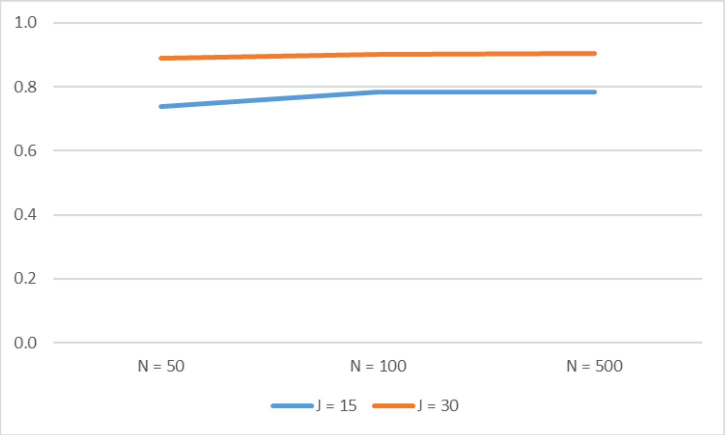


Figure 5. Classification accuracy for the simple dichotomous case.

Note. *N* denotes sample size; *J* denotes test length.

Complex dichotomous case. As shown in Table XIV, the model fit in terms of the PPMC values are satisfactory, as they all larger than .05 or smaller than .95. Not in line with the simple dichotomous case, the biases and the associated *RMSD*s for item parameters fail to show a monotonically decreasing trend but a fluctuating trend instead. The bias for the case of 500 students and 30 items has the smallest value (= .050) but those for the other five cases are all around .06. The *RMSD*s for the first three conditions are slightly lower than the latter three. However, this difference is not large enough (Δ =.010) to alleviate the usage of BN modeling in low-stakes assessments. With respect to person parameters, no pattern of changes in biases and *RMSD*s is shown across conditions. The magnitude of all bias values is relatively small (i.e., around 0.04), and the largest difference (Δ =.015) is between 50 students with 30 items and 500 students with 15 items. The associated *RMSD*s are all around .02. As it turns out, little bias in BN modeling estimation is manifested for both item and person parameters, and these estimates tend to be stable across different conditions of sample sizes and test lengths. The classification accuracies across all conditions are ranged from .54 to .65, showing a moderate agreement between estimated and true classification on student attribute mastery. As shown in Figure 6, the variations among conditions are small, and classification accuracies show little pattern of changes varied by sample sizes and test lengths.

TABLE XIV THE SIMULATION RESULTS FOR THE COMPLEX DICHOTOMOUS CASE

Sample Size	Test Length	Model Fit PPMC	Estimated vs True Item Parameters π		Estimated vs True Person Parameters λ		Classification Accuracy κ
			Absolute -valued Bias	*RMSD*	Absolute -valued Bias	*RMSD*	
50	15	0.366	0.060	0.019	0.041	0.019	0.597
	30	0.292	0.062	0.018	0.051	0.019	0.658
100	15	0.375	0.061	0.016	0.043	0.018	0.636
	30	0.370	0.062	0.026	0.040	0.020	0.608
500	15	0.420	0.057	0.027	0.035	0.021	0.547
	30	0.477	0.050	0.024	0.042	0.025	0.580

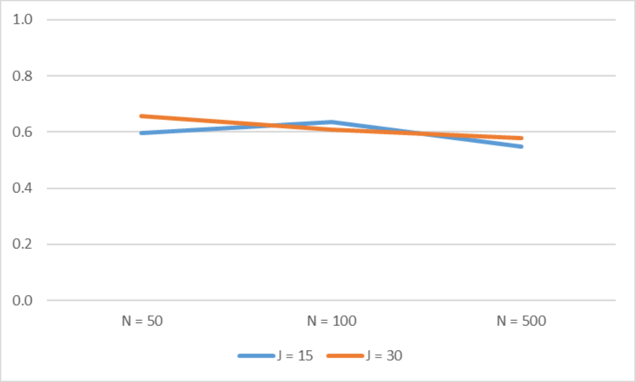


Figure 6. Kappa indices for the complex dichotomous case.

Note. *N* denotes sample size; *J* denotes test length.

Simple polytomous case. Table XV shows the results for the simple polytomous case under the six conditions of sample sizes and test lengths. The PPMC indices for all the conditions are satisfactory, as they all larger than .05 or smaller than .95. The magnitude of biases for item parameter estimation has shown a decreasing trend among conditions. That being said, the condition of 50 students with 15 items has the largest bias (= .052) away from the true item parameter while the condition of 500 students with 30 items has the smallest (= .029). The same tendency is also revealed in the *RMSD*s associated with the biases. These results suggest that for the simple polytomous case, more students and longer test lengths would help to reduce the estimation bias and present more stable parameter estimates in BN modeling. However, the difference is still small (Δ_{bias} = .023, Δ_{RMSD} = .010), especially for a low-stakes assessment. In terms of person parameters, unlike item parameters, no decreasing pattern is found in biases across conditions nor in the associated *RMSD*s. However, the conditions of shorter test length and smaller sample size still show larger bias and higher *RMSD*s. The largest difference in the bias is presented between the case of 50 students and 15 items and the case of 500 students and 15 items (Δ =.008), and the largest difference in *RMSD*s is between the case of 50 students and 15 items and the case of 500 students and 30 items (Δ =.009). These differences are smaller than those of item parameters and can also be regarded as small. In line with the simple dichotomous case, the classification accuracies of the simple polytomous case for the conditions of longer test length (i.e., *J* = 30) regardless of sample sizes are higher than those with 15 items, as shown in Figure 7. All Kappa indices are higher .6, reaching substantial agreement. In general, the application of BN modeling in a simple polytomous case across different conditions of sample sizes and test lengths show little obvious variation in estimation bias and stability.

TABLE XV THE SIMULATION RESULTS FOR THE SIMPLE POLYTOMOUS CASE

Sample Size	Test Length	Model Fit PPMC	Estimated vs True Item Parameters π		Estimated vs True Person Parameters λ		Classification Accuracy κ
			Absolute -valued Bias	*RMSD*	Absolute-valued Bias	*RMSD*	
50	15	0.306	0.052	0.013	0.043	0.020	0.669
	30	0.168	0.048	0.010	0.049	0.018	0.781
100	15	0.274	0.046	0.011	0.045	0.019	0.668
	30	0.151	0.044	0.008	0.050	0.017	0.801
500	15	0.337	0.042	0.010	0.052	0.018	0.672
	30	0.207	0.029	0.003	0.050	0.011	0.810

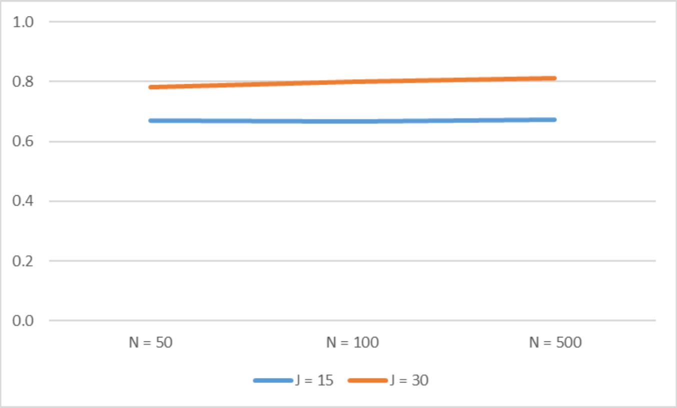


Figure 7. Kappa indices for the simple polytomous case.

Note. *N* denotes sample size; *J* denotes test length.

Complex polytomous case. Table XVI shows the results of the most complex case in this study. All the PPMC indices have met the criteria, as they are all larger than .05 or smaller than .95. They suggest that BN modeling for the complex polytomous case fits the data adequately. The biases of all item parameter estimation have small variations across conditions, with shorter test length showing slightly smaller bias than longer test length. This difference indicates that longer test length for complex Q matrix assessment may involve more complex relationships between attributes and items, thereby more difficult to estimate. The associated *RMSD*s are generally small and are similar to each other, suggesting a stable estimation across conditions. The biases for person parameters are generally small and vary little among different conditions of sample sizes and test lengths (Δ =.006). Their *RMSD*s are very similar to each other with the largest pairwise difference of .002. Further, all the classification accuracies reveal fair to moderate agreement between the estimated and true classification of student attribute mastery. Additionally, Figure 8 presents that the classification accuracies maintain at the similar magnitude regardless of sample sizes and test lengths. It should be noted that the lower classification accuracies in the complex polytomous case compared to previous cases should not preclude the utility of BN modeling in a complex case. One should expect this decrease as in the polytomous attribute case with a complex Q matrix, students got placed into more levels after accounting for the complex relationships between attributes and items, thereby resulting in more variance in the attribute mastery classification than simpler conditions. Classification accuracies are similar across different conditions of sample sizes and test lengths. These results again show that BN modeling for the complex polytomous case is affected little by varying levels of sample size and test length.

TABLE XVI THE SIMULATION RESULTS FOR THE COMPLEX POLYTOMOUS CASE

Sample Size	Test Length	Model Fit	Estimated vs True Item Parameters π		Estimated vs True Person Parameters λ		Classification Accuracy κ
		PPMC	Absolute-valued Bias	*RMSD*	Absolute-valued Bias	*RMSD*	
50	15	0.287	0.051	0.018	0.035	0.019	0.372
	30	0.168	0.056	0.018	0.038	0.019	0.475
100	15	0.274	0.051	0.016	0.034	0.019	0.426
	30	0.151	0.054	0.018	0.037	0.020	0.363
500	15	0.755	0.049	0.018	0.035	0.019	0.373
	30	0.207	0.048	0.016	0.032	0.017	0.382

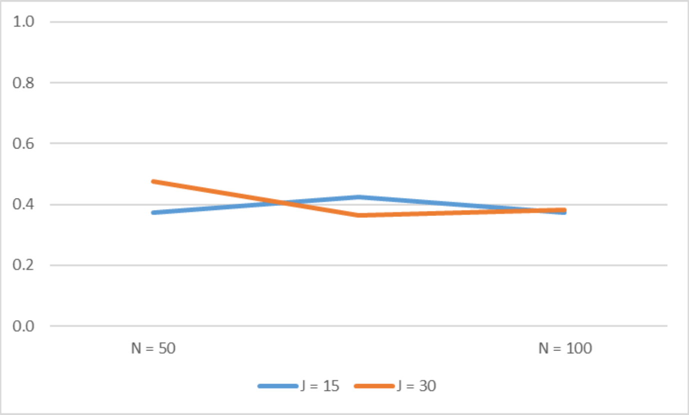


Figure 8. Kappa indices for the complex polytomous case.

Note. *N* denotes sample size; *J* denotes test length.

In summary, little evidence supports that the larger sample size and the longer test length would lead to a significantly better parameter estimation and classification accuracy of attribute mastery across different assessment scenarios when BN is applied. In other words, BN modeling tends to perform adequately for a small sample size of 50 students and a short test length of 15 items and maintains the estimation accuracy and stability with little variations from conditions of a large sample size and a long test length. From another perspective, in most conditions across the four cases, larger sample size and longer test length yield a small increase in estimation accuracy, especially for simple dichotomous and simple polytomous cases. Although small, this variation is consistent with the intuitive assumption that assessments, which measure attributes independently by items, would give models more information on each attribute for parameter estimation compared to a complex case with the same sample size and test length, and therefore would furnish more accurate parameter estimation.

However, considering the interplay among attributes in answering an item correctly, it is necessary to design items to evaluate whether students know how to use multiple attributes simultaneously to solve problems. Additionally, furnishing finer attribute mastery classifications would prepare students and teachers for knowledge remediation and instruction improvement. Although a larger sample size and a longer test length would improve the estimation accuracy, in formative and diagnostic educational assessments, it is usually difficult to find a large sample size (i.e., $N > 500$). To address these practical concerns, BN modeling opens the door for teachers and researchers to diagnostically analyze assessment data with a relatively small sample size and short test length but at the same time yields reasonable parameters estimation accuracy. The results of the first research question show that all the assessment conditions varying by sample size and test length have reached reasonable accuracy with small variations between each other. For this reason, I used a combination of sample size ($N = 100$) and test length ($J = 15$) that is feasible in common practice for the following simulation studies.

4.1.3 Research Question 2a

Based on a stable and optimal selection of sample size and test length (if any) from the first research question, under the different conditions of attribute types and Q matrix structures, how well can the following parameters be recovered: students' mastery profiles, item parameters, person parameters?

This research question entails a comparison of the performance of BN modeling among two kinds of Q complexity and two types of attributes and their parameters under the same sample size and the same test length. From the modeling perspective, the simple Q matrix cases would have more straightforward modeling specification as each item only measures one attribute, and the complex cases tend to be more complicated as they measure students' competence in simultaneously using multiple attributes to solve problems. The polytomous cases need to consider three or more mastery levels in an attribute rather than two levels of mastery and non-mastery in dichotomous cases. Consequently, more complicated assessment contexts may lead to less accurate parameter estimation compared to relatively simpler assessment scenarios. Table XVII presents the results of the four cases. The model fitting results of the four cases show that BN modeling performs well on each case, as they are all larger than .05 or smaller than .95. These findings indicate the flexibility of BN in handling different levels of Q matrix complexity and different types of attributes. With respect to the estimation bias of item parameters across the four cases, the cases of simple Q matrix have smaller biases than those of complex Q matrix, while no apparent difference is found between cases of the two attribute types. The largest difference is between the simple dichotomous case (= .041) and the complex dichotomous case (= .061). However, the biases and the associated *RMSD*s are all small for the four cases. To be specific, the simple dichotomous case has the smallest *RMSD* (= .007) while the complex dichotomous case has the largest (= .016). These results echo the previous assumption that BN modeling for simpler assessment scenarios would give out more accurate estimation. But the small variations among cases also suggest that BN modeling can provide item parameters with acceptable biases for different kinds of Q matrix complexity and different item types.

TABLE XVII SIMULATIONS RESULTS FOR THE FOUR CASES WITH A 15-ITEM TEST AND 100 STUDENTS

Model	Model Fit	Estimated vs True Item Parameters π		Estimated vs True Person Parameters λ		Classification Accuracy κ
	PPMC	Absolute-valued Bias	*RMSD*	Absolute-valued Bias	*RMSD*	
Simple Dichotomous	0.214	0.041	0.007	0.048	0.014	0.784
Complex Dichotomous	0.375	0.061	0.016	0.043	0.018	0.636
Simple Polytomous	0.274	0.046	0.011	0.045	0.019	0.668
Complex Polytomous	0.274	0.050	0.016	0.034	0.019	0.426

In terms of person parameters, the results show that the complex polytomous case has the least bias while the other three cases have similar biases. Different from item parameters, the biases of person parameters for the cases of complex Q matrix tend to be lower than the simple cases. The largest difference is between the case of simple dichotomous (= .048) and the complex polytomous (= .034). Their *RMSD*s are similar to each other. These findings suggest that, although a different pattern is presented for person parameters, the biases and their *RMSD*s are still small. As it turns out, BN modeling can be considered as robust in estimating person parameters across varied assessment scenarios.

Finally, in congruent with the contention that simple cases would yield more accurate model estimation results, the simple dichotomous case shows the highest classification accuracy (= .784) while the complex polytomous case reflects the lowest accuracy (= .426) in classifying student attribute mastery levels. Further, the cases of simple Q matrix have higher Kappa indices than the cases of complex Q matrix. In particular, the complex polytomous case has yielded a moderate agreement between estimated and true classifications while the other three cases have substantial agreements. The modeling complexity, which has impact on the classification accuracy, may lead to the lower agreement of the complex polytomous case. In the complex polytomous modeling, I computed the estimated classification by its average classifications among the 4000 iterations (after 1000 burn-ins). This might be the other reason for the lower agreement of the complex polytomous case. Put more succinctly, if the average classification of an attribute for a person across 4000 iterations is between [0, .7), then the classification is 0; if falls within [.7, 1.4), then the classification is 1; if it fall within [1.4, 2], the classification is 2. By doing so, the classification results might have some biases because they were re-calculated by assuming the average classification results follow a uniform distribution. In sum, although the complex polytomous case has slightly lower classification accuracies compared to other cases, the usage of BN modeling in analyzing assessment data of different Q matrix complexities and different attribute types present acceptable classification accuracy.

4.1.4 <u>Research Question 2b</u>

Based on the same selection of sample size and test length, how well the classification accuracies of student attribute mastery can be recovered with a prior information of no structural relationships, partial structural relationships, full structural relationships, and wrong structural relationships? What cognitive diagnostic information can be provided to students and teachers?

In the previous two research questions, I hypothesized that the true structural relationships among attributes are already known based on the structure in attributes shown in Figure 4. However, in a real-world situation, teachers or researchers may not know the true structural relationships among the latent variables of skills as these skills are unobservable. Therefore, it is necessary to evaluate whether BN modeling can recover the diagnostic attribute classifications when the information about the structure embedded among attributes is unknown. To achieve this goal, I set four conditions of prior information on the structural relationships among attributes: no prior information, partial information, full information, and wrong information. Specifically, under the condition of no prior information, there are no relationships among attributes. In other words, attributes are assumed to be independent in this scenario. Under the partial information condition, I used the structure shown in the left graph of Figure 9. I removed the edge from attribute 2 to 5 and the edges from attributes 2 and 5 to 4 in the original structure (see in Figure 4). For the condition of wrong information, I used the structural relationships shown in the right graph of Figure 9. I reversed the arrow from attribute 2 to 3 and the one from attribute 2 to 5 in the original structure and removed the paths from attributes 2 and 5 to attribute 4.

Using the same combination of 100 students and 15 items, I evaluated the classification accuracy of student attribute mastery under each condition of prior information on the structure in attributes for each kind of Q matrix complexity and attribute type (see in Table XVIII).

TABLE XVIII THE SIMULATION RESULTS OF FOUR CASES UNDER DIFFERENT CONDITIONS OF PRIOR INFORMATION ON STRUCTURE IN ATTRIBUTES

Model	Prior Information on Structure in Attributes			
	No Info	Partial Info	Full Info	Wrong Info
Simple Dichotomous	0.736	0.764	0.784	0.770
Complex Dichotomous	0.494	0.603	0.636	0.609
Simple Polytomous	0.609	0.643	0.668	0.642
Complex Polytomous	0.333	0.413	0.426	0.401

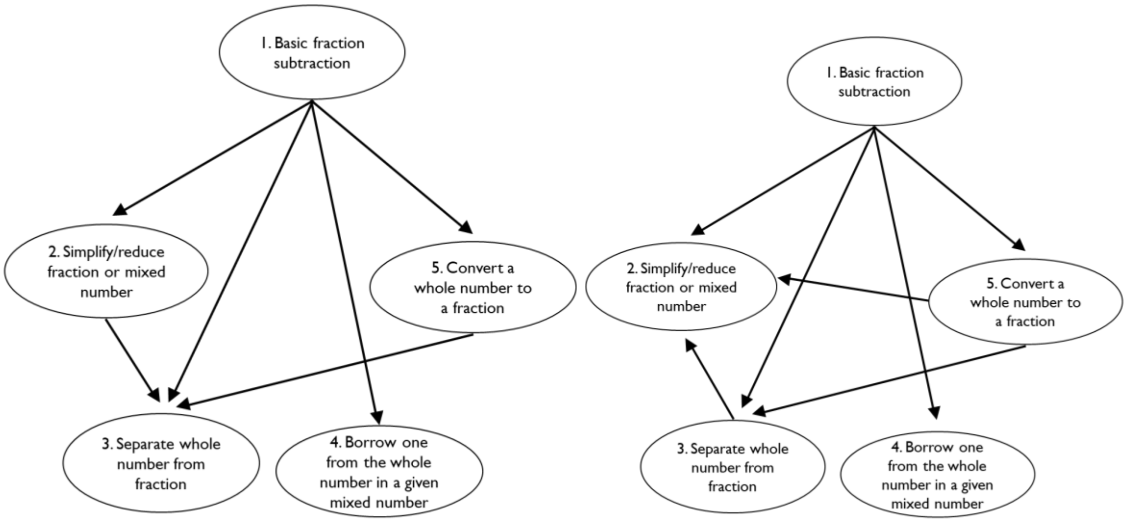

1. Basic fraction subtraction
2. Simplify/reduce fraction or mixed number
5. Convert a whole number to a fraction
3. Separate whole number from fraction
4. Borrow one from the whole number in a given mixed number
1. Basic fraction subtraction
2. Simplify/reduce fraction or mixed number
5. Convert a whole number to a fraction
3. Separate whole number from fraction
4. Borrow one from the whole number in a given mixed number

Figure 9. The partial structure of attributes (Left); The wrong structure of attributes (Right).

Within each condition of prior information, the four cases manifest the pattern that the cases of simple Q matrix would yield higher classification accuracy than the complex cases, and the cases of dichotomous attributes tend to show higher classification accuracy than polytomous attributes as shown in Table XVIII. After removing some paths from the true structural relationships of attributes, the condition of partial information mirrors the situation where teachers or researchers partially understand the difficulty of each attribute, or the learning progression of attributes from some previous studies, but they are unsure about the accurate structure. The results of this scenario indicate that when we fed the model with less information, the classification accuracy tends to be slightly lower than the condition where the complete true relationships were captured in the modeling. According to the rule of thumb for the Kappa statistic, the classification accuracies of the simple dichotomous (κ = .764), complex dichotomous (κ = .603), simple polytomous cases (κ = .643) still reach a substantial agreement, and the complex polytomous case still maintains a moderate agreement (κ = .413). It reveals that BN modeling would still have its diagnostic classification accuracy reach a satisfactory level across different kinds of Q matrices and different attribute types when partial hypotheses on the structure in attributes are provided.

Another condition I evaluated for this research question is when we missed a structural relationship among attributes. In other words, we assume that the acquisition of one attribute is not associated with that of another attribute, and each independently contributes to item responses. The results of this scenario show that the simple dichotomous case (κ = .736) and the simple polytomous case (κ = .609) are smaller than the partial information condition but still maintain the substantial agreement between the estimated and true student attribute mastery status. It is probably because the modeling for the cases of simple Q matrix complexity is free from the impact of interactions between attributes on item responses so that it is impacted less even without prior information on attribute structure. However, for the cases of complex dichotomous (κ = .494) and complex polytomous (κ = .333), the classification accuracies drop to moderate agreement and fair agreement. This discrepancy with the partial information and full information conditions shows that ignoring the prior information on attribute structure may reduce its modeling accuracy in making diagnostic classification for student attribute mastery, especially for the cases of complex Q matrices.

Finally, it is also possible that teachers or researchers may have a wrong assumption about the attribute structure. Teachers and researchers may have some knowledge about which are the basic attributes (attribute 1) and the advanced attributes (attribute 4), but not sure about the dependent relationships among the intermediate attributes (attributes 2, 3, 5). Table XVIII has shown that the classification accuracy for the condition of wrong information tends to be lower than the full information condition, but similar to the case being provided with partial information. These results indicate that although being fed with some wrong information, BN modeling is still able to recover students' attribute mastery to an acceptable level. Note that it is also important for teachers and researchers to have some information on the attribute structure, and it is still worth telling the model with some basic assumptions even they might not be accurate.

For an illustrative purpose, I graphically displayed the structure in attributes and their relationships with items under different scenarios of prior information using Netica (Norsys, 1992-2014) in Figures 10 to 15. In Netica, I modeled the posterior distribution over the proportion of masters for each attribute and the probability of correctness for each item for the complex dichotomous case with no prior information, partial prior information, full prior information, and wrong information. I also used a student's item responses to showcase the relationships between attribute structure and item responses. The model formulated in Netica for the case of full prior information on structure in attributes is given in Figure 10. In the graph, each node is depicted by a bar revealing the probability for the node being in each state, which is also represented numerically as a percentage. They denote the average probability of correctness for each item denoted by *X* and the average proportion of mastery for each attribute denoted by alpha. The probabilities of correctness can be compared to check item difficulties. As can be found from Figure 10, items measuring more attributes tend to be more difficult than items measuring one or two attributes. For example, items 1, 2, and 3 measuring the basic attributes 1 and/or 2 have higher probabilities of correctness than other items. Also, among items 7-11 measuring the same attributes 1, 3, and 4, items 7, 9, 10 are more difficult than items 8 and 11. The proportion of masters for each attribute manifest the difficulty of mastering this attribute. The results show that attributes 1 and 2 are easier ($prob_{\text{master}} > .80$) to master than attribute 5 ($prob_{\text{master}} = .775$), and attributes 3 ($prob_{\text{master}} = .627$) and 4 ($prob_{\text{master}} = .577$) are harder for students to acquire.

Figure 11 shows the joint distribution over the true distribution of students' attribute mastery levels and the distribution of simulated item responses. In comparison with Figure 10, the true joint distribution slightly differs from the estimated distribution of the full information condition. However, this difference does not affect the general distribution of difficulties among attributes, indicating a consistent pattern between the true and the BN estimated results.

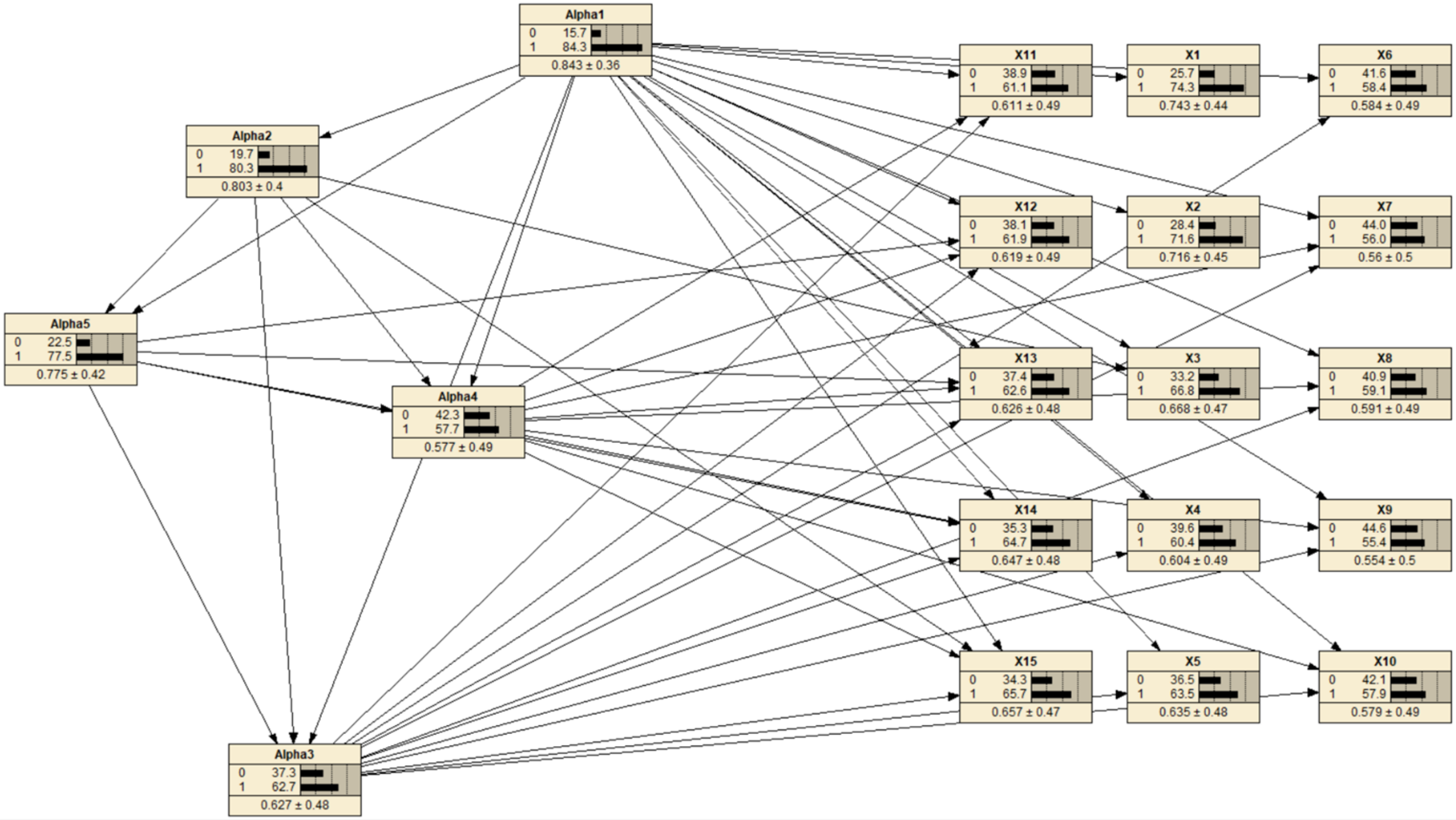


Figure 10. The graphical display of the BN posterior distribution under the condition of full information.

Note. Alpha denotes the mastery classification for each attribute, *X* denotes the items responses to each item.

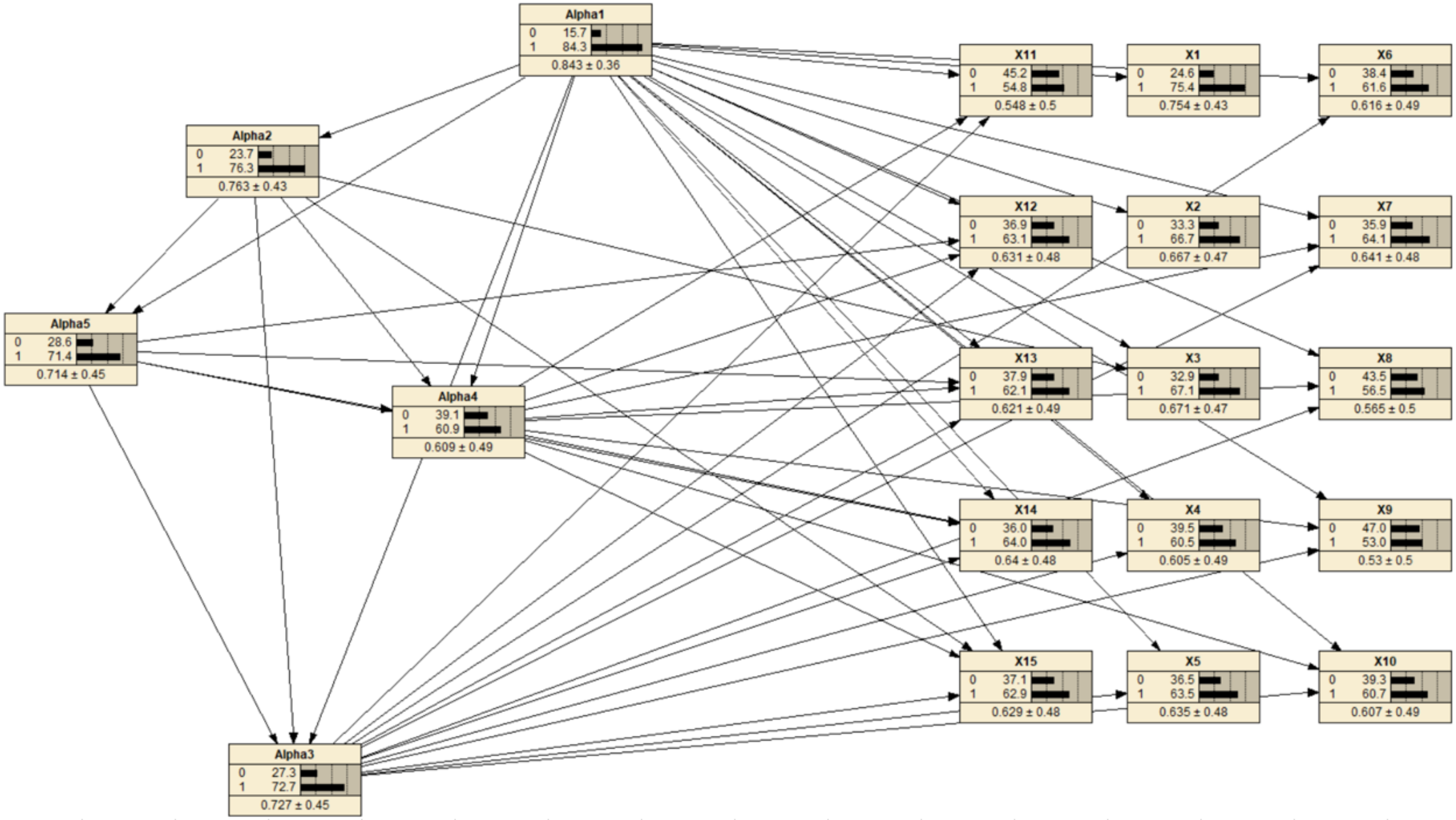


Figure 11. The graphical display of the true joint distribution.

Note. Alpha denotes the mastery classification for each attribute, *X* denotes the items responses to each item.

Further, the flexibility of the BN formulated in Netica allows entries of specific states for each variable in order to observe how the probabilities of other variables are influenced accordingly. For example, based on the scenario where the full information is provided, teachers would like to know how the probabilities of a student's mastery of attributes 3 and 4 and how the probabilities of item correctness are impacted if this student mastered attribute 1 and 2 but not attribute 5. This scenario is depicted in Figure 12. The results indicate that this student have a 25% probability of mastering attribute 3 and a 12.5% probability of mastering attribute 4. In addition, the probabilities of correctness for items (items 7-12) measuring attributes 3, 4, and 5, which this student failed to master, become smaller than the average level shown in Figure 10.

With respect to the partial prior information condition, Figure 13 shows that attribute 5, which lacks a path compared to the true structure, becomes slightly easier to acquire as its proportion of masters turns slightly higher. Attribute 3 tends to become harder as it requires more prerequisite attributes than other attributes. For items 12-14 requiring both attributes 3 and 5, their probabilities of correctness show small difference ($\Delta \approx .04$) compared to the true joint distribution shown in Figure 11. These results suggest that BN modeling with partial information would provide results that deviate little from the truth.

For the wrong information condition shown in Figure 14, it can be found that, similar to the partial information case, the posterior distribution over attributes has some small differences compared to the true joint distribution shown in Figure 11. Specifically, attribute 5 becomes slightly easier while attributes 2 and 3 turn slightly harder. This small difference again shows the capacity of BN modeling in recovering the true classification results to some extent.

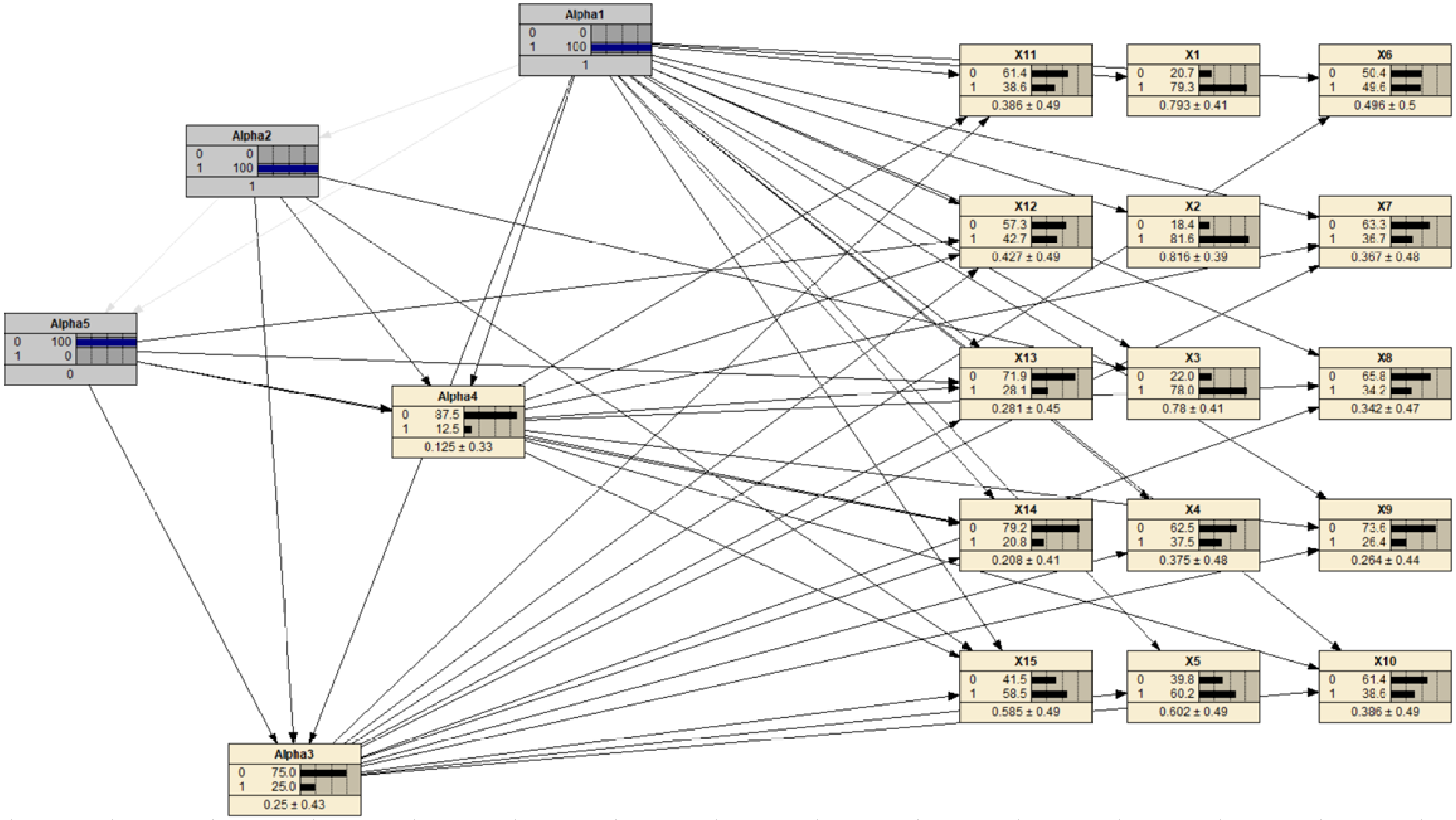


Figure 12. The graphical display of the BN posterior distribution with some attributes fixed with mastery levels.

Note. Alpha denotes the mastery classification for each attribute, *X* denotes the items responses to each item.

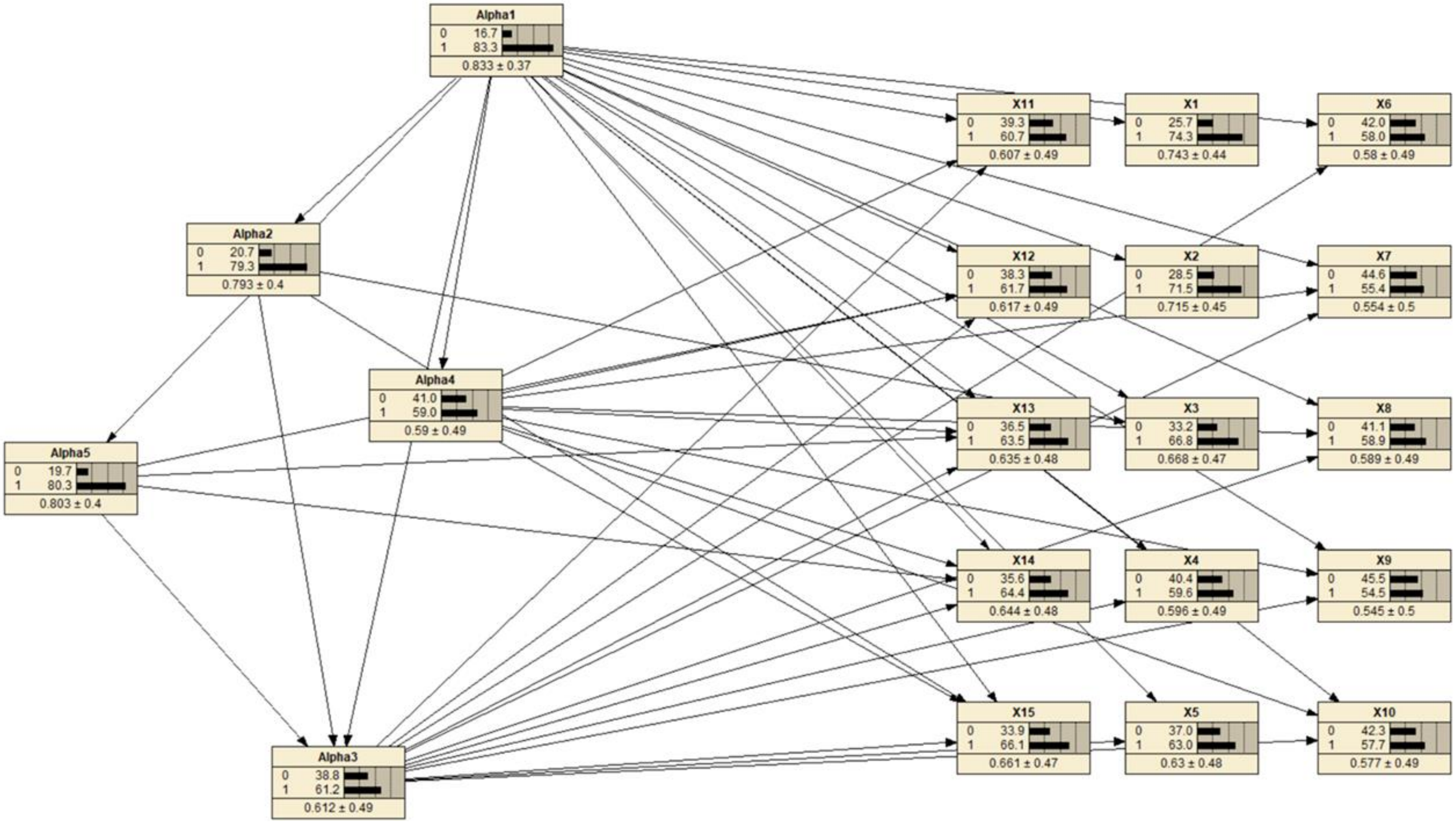


Figure 13. The graphical display of the BN posterior distribution under the partial information condition.

Note. Alpha denotes the mastery classification for each attribute, *X* denotes the items responses to each item.

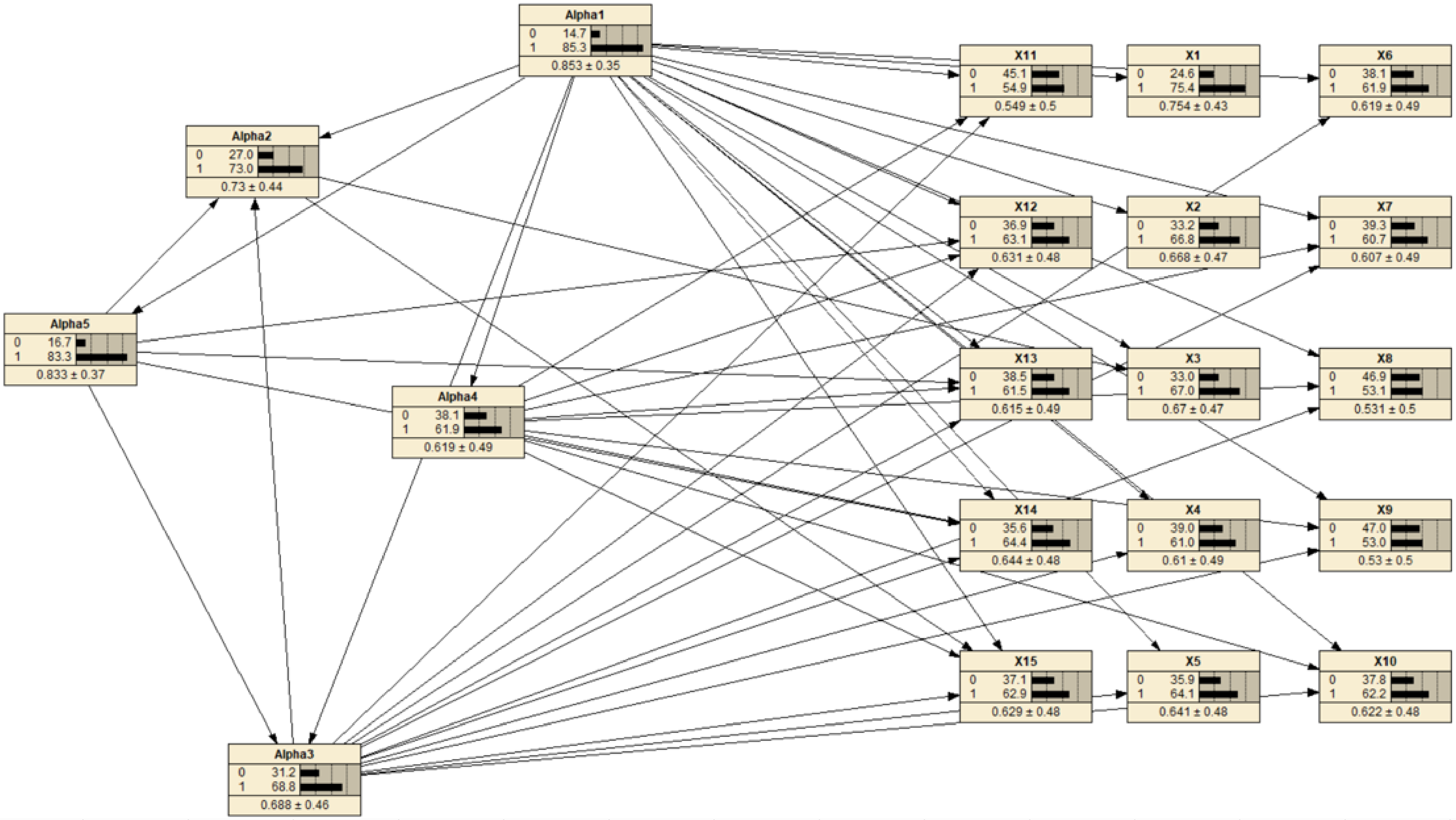


Figure 14. The graphical display of the BN posterior distribution under the wrong information condition.

Note. Alpha denotes the mastery classification for each attribute, *X* denotes the items responses to each item.

The modeling for the third condition considered no structure in attributes and hypothesized that they independently affect students' responses to items as shown in Figure 15. Not surprisingly, some difference shows in the proportion of masters for each attribute when compared to the true joint distribution shown in Figure 11. The mastery probabilities for attributes 2, 3 and 5 become easier while attribute 3 becomes harder. Regarding the probabilities of correctness for items, more disparity is found than the full or partial information conditions. The items measuring multiple attributes have a difference around .04 in their average probabilities of correctness, higher than the difference found in the partial and wrong information conditions. These results suggest that analyzing item responses without any prior information on the structure in attributes would affect the estimated proportions of mastery in attributes and the estimated probabilities of correctness for items.

I used Examinee 85 as a case study for illustration. Table XIX lists the response vector for this student. I then entered this student's responses into the Netica BN model built with the full information. Figure 16 shows that examinee 85 performed well on the items that require a simultaneous usage of attributes 1, 2, and 5, but struggled with the items that require attributes 4 and 5. This student is very likely to possess attributes 1, 2, and 5 as their mastery probabilities are larger than or close to 90%, and this student is nearly certain to fail to acquire attribute 3 (*prob* = 15.8%) and may not possess attribute 4 with a mastery probability of 33.5%. This finding is consistent with the true mastery status of this student. This BN demonstration allows an interpretation of student attribute mastery status given their item responses.

TABLE XIX Q MATRIX AND ITEM RESPONSES FOR EXAMINEE 85

Item ID	Attributes					Item responses
	1	2	3	4	5	
1	1					1
2	1					1
3	1	1				1
4	1		1			0
5	1		1			1
6	1		1			0
7	1		1	1		1
8	1		1	1		0
9	1		1	1		0
10	1		1	1		1
11	1		1	1		0
12	1		1	1	1	0
13	1		1	1	1	1
14	1		1	1	1	1
15	1	1	1	1		0

Taken together, these results demonstrate that less information on the structure in attributes would yield less satisfactory model fitting and less accurate diagnostic classification results when the structure in attributes truly exists. However, the application of BN modeling can maintain the classification accuracies at a satisfactory level for conditions where at least some information or assumptions are given. Further, the graphical features of BN modeling allow a straightforward display of posterior distribution over attributes and items. It can also unfold diagnostic results for each student on their mastery of each attribute when item responses are observed.

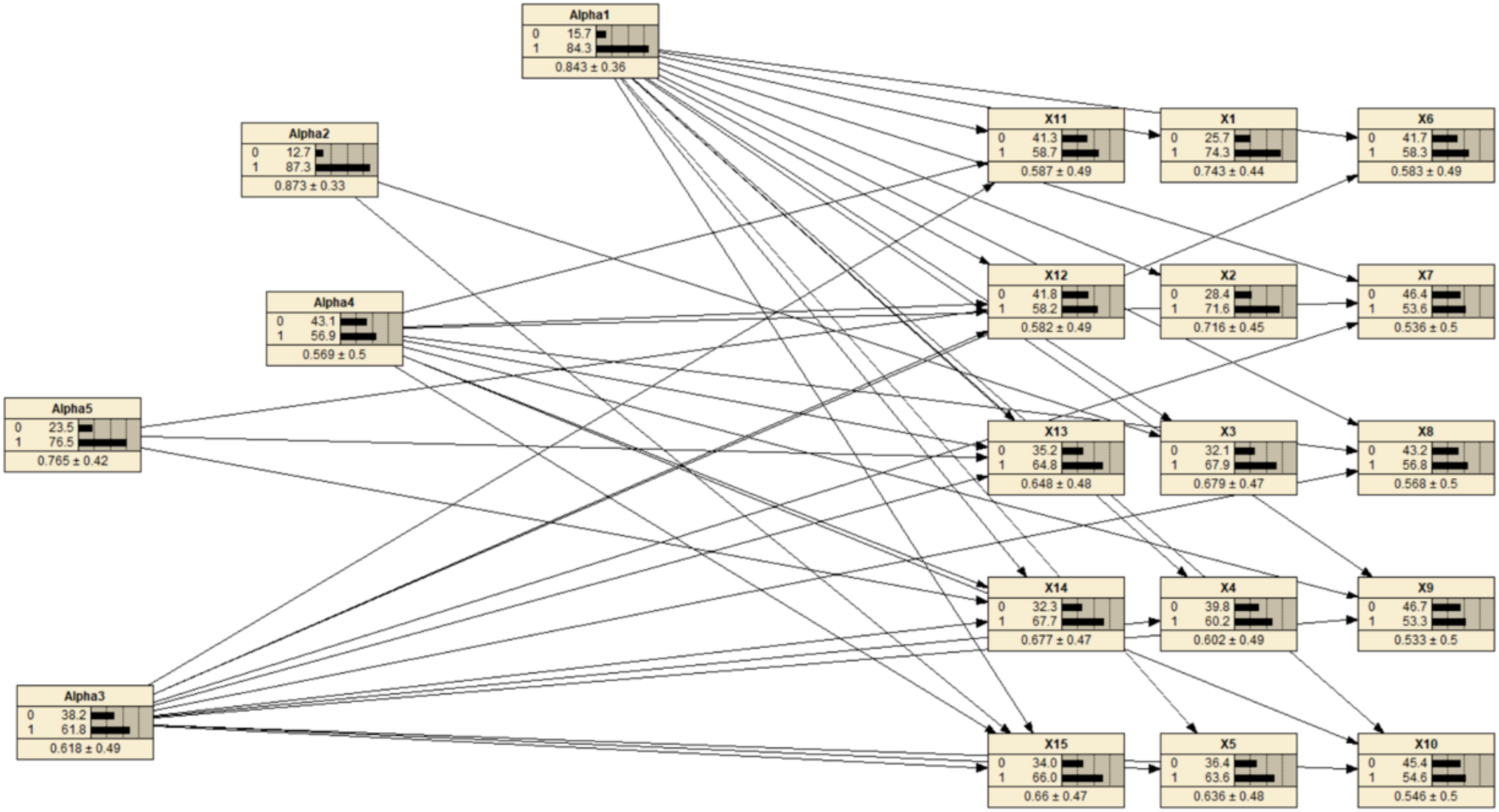


Figure 15. The graphical display of the BN posterior distribution under the no information condition.

Note. Alpha denotes the mastery classification for each attribute, *X* denotes the items responses to each item.

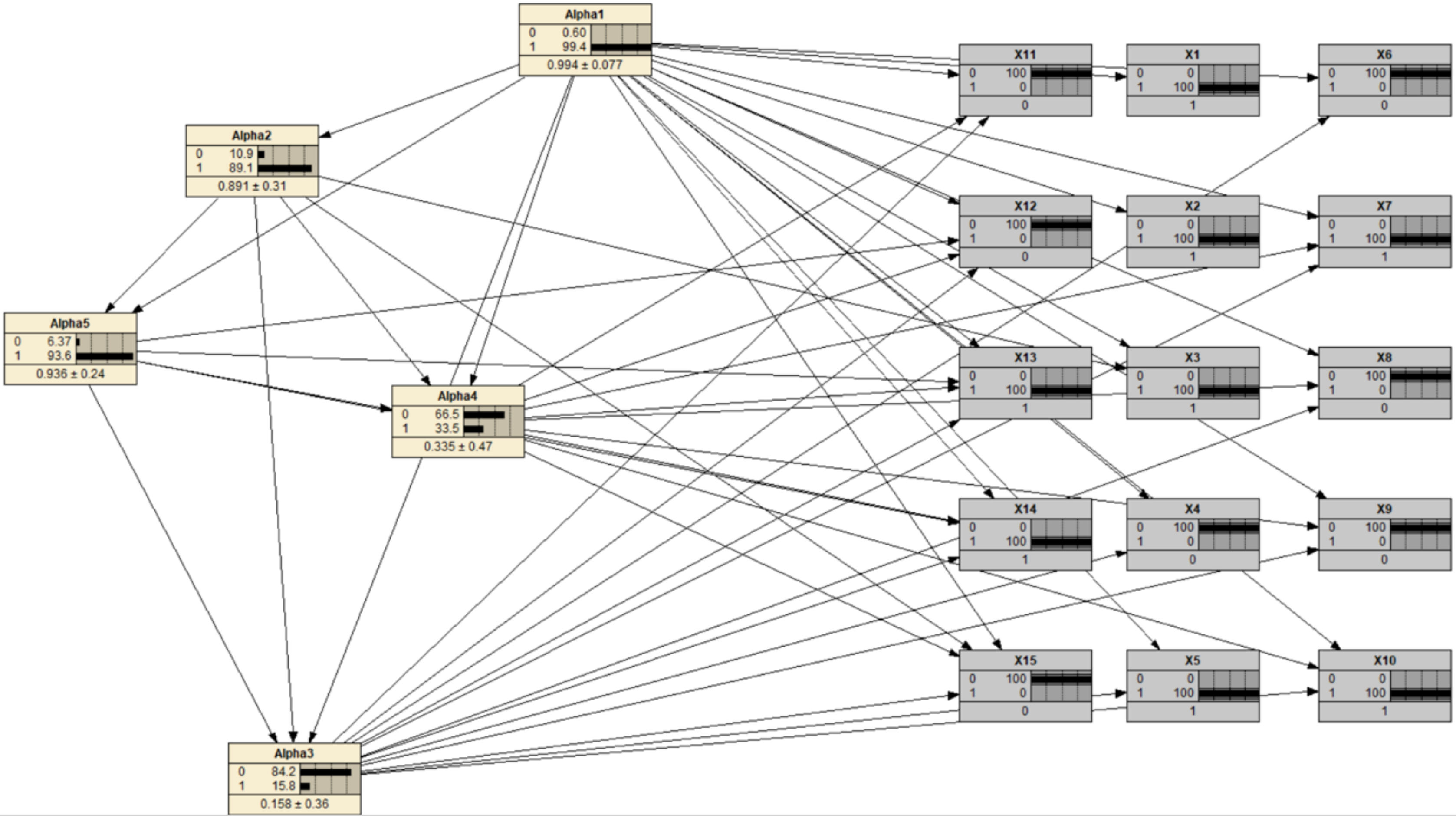


Figure 16. The graphical display of the BN posterior distribution under the full information condition with the Examinee 85 item responses fixed.

Note. Alpha denotes the mastery classification for each attribute, *X* denotes the items responses to each item.

4.1.5 Research Question 2c

Based on the same selection of sample size and test length as previous questions, how are the models' fitting and accuracies different between the CDM and the BN approach?

As another modeling approach for diagnostic assessments, CDMs have gained attention in both research and practice. As demonstrated in Section 2.3.3, the similarities between CDM and BN warrant the need to compare the two related but distinct approaches which are both built under the latent class modeling framework. In the practical application, the classification results of CDMs are sometimes refrained from rendering more useful information due to its compensatory rules. In other words, most commonly used CDMs (i.e., DINA, DINO) classify students into two groups either who have mastered all the measured attributes and who have not or who have mastered at least one measured attribute and who have not mastered any. By doing so, many information about students who are in different mastery levels (e.g., having mastered 2 out of 3 measured attributes) is lost in such modeling rules. Further, it may contribute less to the requests from teachers or researchers who would like to have finer-grained feedback on student mastery levels. Later, a saturated CDM, GDINA (De La Torre, 2011), which can consider all mastery levels, releases this constraint. When comparing CDM with BN, it would be fair to use GDINA as both are saturated approaches and both can handle structure in attributes. To evaluate the performance of both approaches in analyzing formative diagnostic assessments, I compared the model fit indices and the classification accuracies between GDINA and BN. I also compared their performance under different prior information conditions. Similar to the previous research questions, 20 datasets were replicated for each comparison.

Note that during the modeling process, I found some limitations of GDINA. First, although the GDINA model can consider the structural relationships among attributes, it takes these relationships as prerequisites rather than correlations. In other words, the structure shown in Figure 4 is interpreted as mastering attribute 1 is required to master attribute 2, mastering attributes 1 and 2 is required to master attribute 5, mastering attributes 1, 2, 5 is required to master attributes 3 and 4, respectively. In this case, students who have mastered attributes without mastering the prerequisite attributes (e.g., students have not mastered attributes 1 and 2 but mastered attribute 5) are considered as impossible and therefore are excluded in classifications. By doing so, the input of structural relationships among attributes reduces latent classes by removing mastery profiles incongruent with the prerequisite rule. Second, I also found that although the GDINA model can handle polytomous attributes, they fail to simultaneously accommodate the structural relationships among polytomous attributes. It is also true for other commonly used CDMs. Therefore, no analysis was conducted in this study for polytomous-attribute GDINA under the conditions of partial, wrong or full information on the structural relationships among polytomous attributes.

Table XX presents the model fit indices and the classification accuracies for both models under the four conditions of prior information when available. As demonstrated in Section 3.3.5.1, deviance is used to evaluate the model fit of the two models. When assuming attributes are independent without any structural relationships (i.e., the no information condition), it can be found that the deviances of the BN model for the simple dichotomous, complex dichotomous, simple polytomous cases are smaller than those of the GDINA model, suggesting that BN yields better model fit than GDINA on these cases. While for the complex polytomous case, the GDINA model has slightly better model fit than the BN model. In terms of the attribute mastery classification accuracy, Kappa statistic of the two models for the simple dichotomous case are very similar and maintain substantial agreement. While for the complex dichotomous case, the classification accuracy for BN modeling (κ = .494) is higher than the GDINA modeling (κ = .385), with the former having moderate agreement and the latter having fair agreement. In the simple polytomous case, GDINA has a slightly better classification accuracy than BN, both having substantial agreement. Although GDINA has a better model fit in the complex polytomous case, its classification accuracy (κ = .306) is lower than BN (κ = .333), both having fair classification accuracy. I also evaluated the classification agreement between BN and GDINA. As can be seen, they have higher agreement in the simple Q matrix cases than the complex ones, and the dichotomous cases have higher agreement than the polytomous cases. These results reveal that for assessments in which attributes have relationships with each other in reality, but no prior information is provided to the model, BN generally performs better than GDINA in terms of a better model fit and a higher classification accuracy. The classification consistency results show that their classification results are congruent with each other.

TABLE XX THE COMPARISON BETWEEN BN AND GDINA UNDER DIFFERENT CONDITIONS OF PRIOR INFORMATION

	Conditions of prior information																			
Condition	No information					Partial information					Full information					Wrong information				
	Model fit: Deviance		Mastery Profile Accuracy			Model fit: Deviance		Mastery Profile Accuracy			Model fit: Deviance		Mastery Profile Accuracy			Model fit: Deviance		Mastery Profile Accuracy		
	BN	GDINA	BN	GDINA	BN vs. GDINA	BN	GDINA	BN	GDINA	BN vs. GDINA	BN	GDINA	BN	GDINA	BN vs. GDINA	BN	GDINA	BN	GDINA	BN vs. GDINA
Simple Dichotomous	1459	1744	0.74	0.73	0.79	1474	1737	0.76	0.71	0.82	1512	1779	0.78	0.73	0.86	1489	1759	0.77	0.73	0.84
Complex Dichotomous	1611	1660	0.49	0.39	0.51	1630	1709	0.60	0.45	0.59	1630	1718	0.64	0.43	0.62	1617	1701	0.61	0.49	0.65
Simple Polytomous	1491	1705	0.61	0.63	0.80	1451	NA	0.64	NA	NA	1505	NA	0.67	NA	NA	1457	NA	0.64	NA	NA
Complex Polytomous	1654	1628	0.33	0.31	0.48	1687	NA	0.41	NA	NA	1672	NA	0.43	NA	NA	1681	NA	0.40	NA	NA

For the condition of providing partial information on the structure in attributes, only the dichotomous cases are compared because the polytomous attribute GDINA is not compatible with attribute structure. The deviances of the two dichotomous cases reveal that BN has better model fit than GDINA. Further, the classification accuracies of BN are also higher than GDINA. Specifically, BN has produced substantial agreement on the two dichotomous cases (κ_{simple} = .764, $\kappa_{complex}$ = .603) while GDINA has maintained substantial agreement on the simple dichotomous case (κ = .710) but dropped to moderate agreement on the simple polytomous case (κ = .447). Likewise, the agreement between the two models has reached the perfect level in the simple dichotomous case (κ = .818) but has maintained the moderate level in the complex dichotomous case (κ = .589). Similar to BN, the classification accuracy of GDINA is higher for the partial information condition than the no information condition. Again, these results support that BN performs better than GDINA. The classification consistency is also high between the two approaches.

With respect to the full information condition, BN still has better model fit than GDINA in the two dichotomous cases. The classification accuracies of both models under this condition are higher than the conditions with partial or no prior information except that the classification accuracy of GDINA for the complex dichotomous case is slightly lower than the condition of partial information. It may suggest that providing more information on attribute structure may contribute little to the classification accuracy for the GDINA modeling. Regarding the model-wise comparison, the Kappa statistic for BN modeling (κ = .784) is higher than the GDINA modeling (κ = .727) for the simple dichotomous case and both have reached the substantial agreement with the true attribute classification. BN, maintaining the substantial agreement, has a higher classification accuracy (κ = .636) than GDINA (κ = .432) in the complex dichotomous case, which has a moderate agreement. The two approaches still maintain high classification consistency.

Finally, regarding the condition where wrong information is provided, the deviances of BN are lower than those of GDINA, indicating a better model fitting. The classification accuracies of BN are also higher than GDINA, in which the accuracies associated with the complex Q matrix case are lower than those of the simple case. The classification consistency between the two approaches is still high.

In summary, although not all comparisons between BN and GDINA under each condition can be conducted due to the limitations of GDINA in a simultaneous compatibility of polytomous attributes and the structure in attributes, the current comparisons exhibit the better flexibility of the BN approach in terms of defining the attribute structure and its latent groups. With different levels of prior information on attribute structure, BN performs better than GDINA in most cases in term of model fit and diagnostic classification results. This finding highlights the capacity and the flexibility of BN in classifying students into true latent mastery classes when compared to GDINA as BN can well consider the structural relationships among attributes and therefore can yield results more representative of true attribute classifications.

4.2 Real Data Analysis

This section answers the third research question, "Based on two existing data sets, how is the effectiveness of the BN approach based on MCMC estimation in analyzing real test data? What cognitive diagnosis can be provided by the BN approach?" and shows the results of the two real data analyses.

4.2.1 MCMC Convergence

To determine the number of iterations to use in analyzing the two datasets, I used the automatic function in the JAGS program to confirm the number of iterations required to converge. It turned out that the TIMSS data needs 40000 iterations and the polytomous attribute data needs 6000 iterations to converge. I then set the iterations accordingly. As same as the simulation study, I used the first 25% of iterations as burn-ins, two chains, and the thinning interval as 1 (i.e., without thinning). To double-check if all parameters have reached convergence, I investigated if the $\hat{R}$ value associated with each parameter is lower than 1.2. The results suggest that all the parameters converged well and are ready for the next-step analysis.

4.2.2 TIMSS Data Analysis

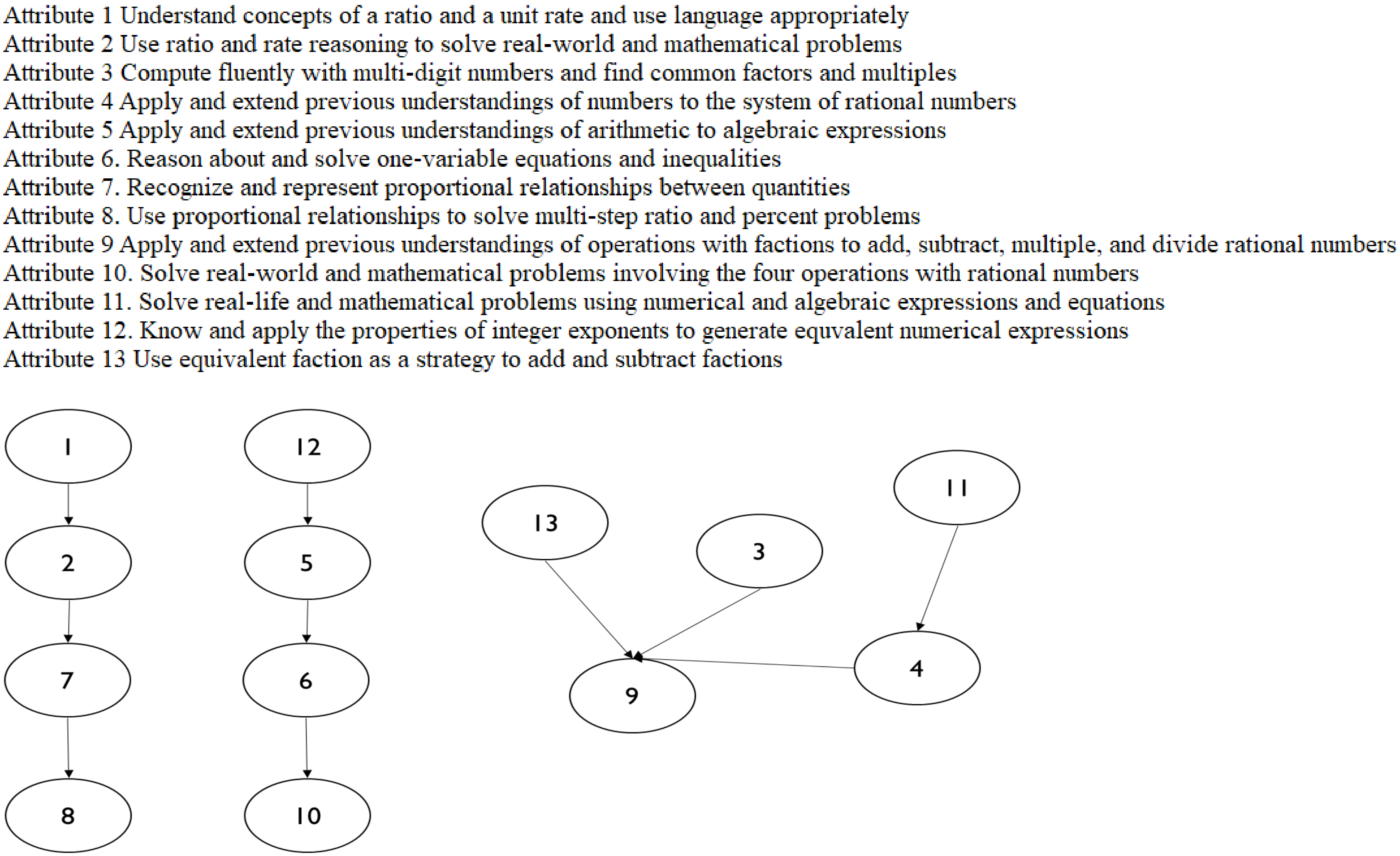
The dichotomous attribute test data, used in Su et al. (2013), came from the United States sample of the Trends in International Mathematics and Science Study (TIMSS) 2003 Grade 8 Mathematics test carried out by the International Association for the Evaluation of Educational Achievement. The data in this study include 757 examinees and their responses to 21 dichotomously scored items, including 19 multiple-choice questions and two constructed response questions. The test measured 15 attributes of math knowledge. Table XXI specifies the Q matrix used in this test, and Figure 23 shows the attributes and their hierarchical relationships used in the data analysis.

Priors. I selected the starting values for parameters based on some assumptions about the attribute mastery and the relationships between attributes and items. Table XXII lists all the priors for each parameter. In terms of priors for person parameters, based on the structural relationships among attributes, the first layer of attributes (i.e., attributes 1, 13, 3, 11, 14), which are not dependent on other attributes, are considered as basic attributes. In this case, these attributes were assumed to be mastered by around 80% of the target population and therefore the associated hyperparameters λ were sampled from beta (8,2). The second layer of attributes (including attributes 2, 4, 5) are dependent on one of the first layer attributes so their probabilities of mastery based on the mastery status of the prerequisite attribute was assumed to .295 by sampling from Beta (7,3) for those who fail master the prerequisite attribute and .705 by sampling from Beta (3,7) for those who have mastered the prerequisite attribute. The third layer of attributes dependent on one of the second layer attributes are considered as harder to master and have probabilities of mastery of .404 by sampling from Beta (4,6) for students with a lack of the prerequisite attribute and .594 by sampling from Beta (6,4). For attribute 9 which depends on three prerequisite attributes, I used the beta distributions of (2,8), (4,6), (6,4), (8,2) to represent the probabilities of mastery on attribute 9 when students mastered none, one, two and all of the prerequisite attributes. The corresponding probabilities sampled from these distributions are .196, .404, .594, .800.

TABLE XXI Q MATRIX FOR THE EIGHTH GRADE TIMSS 2003 MATHEMATICS TEST

Item Number	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13
1	1	0	0	0	0	0	0	0	0	0	1	0	1
2	0	0	0	0	0	1	0	0	0	0	0	0	0
3	0	1	0	0	0	0	1	0	0	0	0	0	0
4	0	0	0	1	0	0	0	0	1	0	0	0	0
5	0	0	0	0	0	1	0	0	0	1	0	1	0
6	0	0	0	0	0	1	1	0	0	0	0	0	0
7	1	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	1	0	0	0	1	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	1	0	0
10	0	0	0	0	0	1	0	0	0	0	0	0	0
11	0	1	0	0	0	0	0	1	0	0	0	0	0
12	0	0	1	0	0	0	0	0	0	0	0	0	1
13	0	0	0	0	1	0	0	0	0	0	0	0	0
14	0	0	0	0	0	1	0	0	0	0	0	0	0
15	0	1	0	0	0	0	0	0	0	0	0	1	0
16	0	0	0	0	1	0	0	0	0	0	0	0	0
17	0	0	0	1	0	0	0	0	0	0	0	0	0
18	0	0	1	0	0	0	0	0	1	0	1	0	1
19	0	1	0	0	0	0	0	0	0	0	0	0	0
20	1	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	1	0	0	0	0	0	0	0	0
22	0	1	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	1	0	0	0	0	1	0	0	0	0

Note. This Q matrix is described by Su et al. (2013)



Attribute 1 Understand concepts of a ratio and a unit rate and use language appropriately
Attribute 2 Use ratio and rate reasoning to solve real-world and mathematical problems
Attribute 3 Compute fluently with multi-digit numbers and find common factors and multiples
Attribute 4 Apply and extend previous understandings of numbers to the system of rational numbers
Attribute 5 Apply and extend previous understandings of arithmetic to algebraic expressions
Attribute 6. Reason about and solve one-variable equations and inequalities
Attribute 7. Recognize and represent proportional relationships between quantities
Attribute 8. Use proportional relationships to solve multi-step ratio and percent problems
Attribute 9 Apply and extend previous understandings of operations with factions to add, subtract, multiple, and divide rational numbers
Attribute 10. Solve real-world and mathematical problems involving the four operations with rational numbers
Attribute 11. Solve real-life and mathematical problems using numerical and algebraic expressions and equations
Attribute 12. Know and apply the properties of integer exponents to generate equvalent numerical expressions
Attribute 13 Use equivalent faction as a strategy to add and subtract factions

Figure 17. Upper: the attributes used in the eighth grade TIMSS 2003 mathematics test described by Su et al. (2013); Bottom: the structural relationships among attributes described by Su et al. (2013).

TABLE XXII PARAMETER VALUES ESTIMATED FOR TIMSS DATA

Parameter	Mean	SD	2.50%	Median	97.50%	Priors	Rhat
lambda_1	0.493	0.037	0.423	0.492	0.565	0.800	1.002
lambda_2[0]	0.085	0.028	0.038	0.082	0.146	0.295	1.002
lambda_2[1]	0.917	0.031	0.852	0.920	0.965	0.705	1.004
lambda_3	0.675	0.104	0.484	0.674	0.888	0.800	1.011
lambda_4[0]	0.378	0.141	0.155	0.360	0.714	0.295	1.022
lambda_4[1]	0.764	0.130	0.191	0.787	0.886	0.705	1.002
lambda_5[0]	0.268	0.069	0.141	0.265	0.411	0.295	1.002
lambda_5[1]	0.680	0.064	0.557	0.679	0.805	0.705	1.001
lambda_6[0]	0.176	0.052	0.084	0.173	0.285	0.404	1.004
lambda_6[1]	0.846	0.043	0.757	0.849	0.923	0.594	1.001
lambda_7[0]	0.442	0.134	0.185	0.447	0.682	0.404	1.004
lambda_7[1]	0.836	0.070	0.668	0.848	0.939	0.594	1.060
lambda_8[0]	0.609	0.128	0.349	0.614	0.846	0.404	1.001
lambda_8[1]	0.543	0.128	0.272	0.553	0.771	0.594	1.002
lambda_9[0]	0.183	0.112	0.025	0.163	0.446	0.196	1.002
lambda_9[1]	0.235	0.099	0.073	0.225	0.452	0.404	1.001
lambda_9[2]	0.267	0.133	0.127	0.243	0.801	0.594	1.006
lambda_9[3]	0.945	0.037	0.856	0.952	0.993	0.800	1.001
lambda_10[0]	0.521	0.144	0.242	0.518	0.800	0.404	1.001
lambda_10[1]	0.509	0.145	0.218	0.514	0.771	0.594	1.002
lambda_11	0.885	0.070	0.722	0.898	0.990	0.800	1.001
lambda_12	0.648	0.079	0.495	0.648	0.804	0.800	1.001
lambda_13	0.625	0.115	0.427	0.613	0.858	0.800	1.011
pai1[1,0]	0.167	0.100	0.025	0.150	0.404	0.196	1.001
pai1[1,1]	0.152	0.059	0.052	0.147	0.280	0.404	1.001
pai1[1,2]	0.602	0.143	0.306	0.624	0.835	0.594	1.001
pai1[1,3]	0.978	0.015	0.943	0.980	0.997	0.800	1.001
pai2[2,0]	0.521	0.033	0.454	0.522	0.585	0.196	1.001
pai2[2,1]	0.889	0.023	0.842	0.889	0.932	0.800	1.001
pai3[3,0]	0.136	0.048	0.048	0.135	0.236	0.196	1.001
pai3[3,1]	0.325	0.101	0.178	0.306	0.581	0.503	1.001
pai3[3,2]	0.828	0.045	0.746	0.825	0.924	0.800	1.001
pai4[4,0]	0.304	0.083	0.136	0.305	0.461	0.196	1.001
pai4[4,1]	0.535	0.052	0.439	0.533	0.645	0.503	1.002
pai4[4,2]	0.753	0.040	0.687	0.751	0.847	0.800	1.001
pai5[5,0]	0.146	0.088	0.021	0.132	0.353	0.196	1.001
pai5[5,1]	0.316	0.104	0.130	0.310	0.537	0.404	1.001
pai5[5,2]	0.750	0.087	0.556	0.758	0.896	0.594	1.001
pai5[5,3]	0.878	0.070	0.718	0.888	0.982	0.800	1.001
pai6[6,0]	0.379	0.119	0.114	0.396	0.565	0.196	1.001
pai6[6,1]	0.865	0.043	0.774	0.869	0.940	0.503	1.001
pai6[6,2]	0.977	0.013	0.946	0.979	0.996	0.800	1.001
pai7[7,0]	0.558	0.034	0.491	0.560	0.619	0.196	1.001

pai7[7,1]	0.924	0.019	0.884	0.925	0.958	0.800	1.001
pai8[8,0]	0.175	0.045	0.085	0.176	0.262	0.196	1.001
pai8[8,1]	0.787	0.052	0.679	0.789	0.881	0.503	1.003
pai8[8,2]	0.962	0.017	0.925	0.964	0.990	0.800	1.021
pai9[9,0]	0.338	0.164	0.057	0.333	0.643	0.196	1.001
pai9[9,1]	0.837	0.025	0.787	0.837	0.884	0.800	1.001
pai10[10,0]	0.152	0.026	0.101	0.152	0.205	0.196	1.002
pai10[10,1]	0.554	0.032	0.492	0.553	0.618	0.800	1.001
pai11[11,0]	0.242	0.106	0.052	0.242	0.448	0.196	1.001
pai11[11,1]	0.712	0.078	0.555	0.715	0.855	0.503	1.001
pai11[11,2]	0.904	0.052	0.784	0.910	0.984	0.800	1.001
pai12[12,0]	0.171	0.101	0.025	0.155	0.407	0.196	1.001
pai12[12,1]	0.223	0.059	0.110	0.222	0.341	0.503	1.001
pai12[12,2]	0.940	0.032	0.869	0.944	0.990	0.800	1.001
pai13[13,0]	0.274	0.029	0.218	0.274	0.333	0.196	1.001
pai13[13,1]	0.670	0.031	0.609	0.670	0.733	0.800	1.001
pai14[14,0]	0.253	0.034	0.187	0.254	0.319	0.196	1.001
pai14[14,1]	0.721	0.030	0.662	0.721	0.781	0.800	1.001
pai15[15,0]	0.202	0.073	0.053	0.206	0.337	0.196	1.001
pai15[15,1]	0.508	0.063	0.389	0.507	0.634	0.503	1.001
pai15[15,2]	0.916	0.034	0.847	0.918	0.978	0.800	1.012
pai16[16,0]	0.262	0.039	0.185	0.263	0.337	0.196	1.004
pai16[16,1]	0.891	0.027	0.836	0.891	0.944	0.800	1.002
pai17[17,0]	0.106	0.076	0.016	0.091	0.376	0.196	1.003
pai17[17,1]	0.621	0.059	0.548	0.613	0.828	0.800	1.001
pai18[18,0]	0.187	0.115	0.026	0.167	0.459	0.196	1.001
pai18[18,1]	0.226	0.091	0.077	0.217	0.432	0.352	1.001
pai18[18,2]	0.220	0.046	0.131	0.219	0.314	0.503	1.004
pai18[18,3]	0.263	0.076	0.130	0.258	0.425	0.643	1.002
pai18[18,4]	0.702	0.046	0.614	0.701	0.795	0.800	1.003
pai19[19,0]	0.060	0.017	0.030	0.059	0.095	0.196	1.002
pai19[19,1]	0.422	0.031	0.365	0.422	0.483	0.800	1.002
pai20[20,0]	0.021	0.010	0.005	0.020	0.043	0.196	1.001
pai20[20,1]	0.305	0.030	0.250	0.305	0.366	0.800	1.003
pai21[21,0]	0.154	0.023	0.110	0.153	0.201	0.196	1.001
pai21[21,1]	0.354	0.027	0.303	0.353	0.408	0.800	1.033
pai22[22,0]	0.492	0.031	0.432	0.493	0.551	0.196	1.001
pai22[22,1]	0.917	0.021	0.874	0.918	0.955	0.800	1.001
pai23[23,0]	0.167	0.063	0.052	0.165	0.297	0.196	1.001
pai23[23,1]	0.244	0.126	0.129	0.220	0.792	0.503	1.001
pai23[23,2]	0.913	0.034	0.842	0.916	0.970	0.800	1.002

Note. lambda [a] denotes the λ parameter for the condition of *a* mastered prerequisite attributes, where 0 denotes none prerequisite attribute is mastered, 1 denotes one prerequisite attribute is mastered, etc. pai[b, a] denotes the π parameter of Item b when *a* measured attributes are mastered, where 0 denotes none of the measured attributes are mastered, 1 denotes one of the measured attributes is mastered, etc.

Regarding priors for item parameters, I classified them into four groups based on the number of attributes they measure. First, for the items measuring only one attribute, the probabilities of correctness for students who have and have not mastered the measured attribute were assumed to follow Beta (8,2) and Beta (2,8), respectively. Items measuring two attributes were hypothesized to have probabilities of correctness for students who have mastered none, one, two measured attributes following Beta (2,8), Beta (5,5), Beta (8,2), respectively. Items measuring three attributes were assigned to follow Beta (2,8), Beta (4,6), Beta (6,4), Beta (8,2) for their probabilities of correctness for student who have mastered none, one, two and three measured attributes, respectively. Finally, items measuring four attributes were assigned to follow Beta (2,8), Beta (3.5,6.5), Beta (5,5), Beta (6.5,3.5), Beta (8,2) for their probabilities of correctness for student who have mastered none, one, two, three and four measured attributes, respectively.

Parameter values. Table XXII summarizes the estimated person and item parameters, their interquartile range, their $\hat{R}$ statistics by taking an average from 30000 iterations following 10000 burn-ins, and the abovementioned priors. All parameters have $\hat{R}$ statistics smaller than 1.2, suggesting the estimation of these parameters has reached convergence.

Person parameters. Person parameters, denoted by λ (lambda), represent the probabilities of mastering an attribute to a certain level for a student given their mastery status of prerequisite attributes. As shown in Table XXII, the probability of mastering attribute 1 (i.e., Understand concepts of a ratio and a unit rate and use language appropriately) is .493. With the probability of mastering attribute 1 at the classification borderline of .5, this attribute is not good at dichotomously classifying students into mastery and non-mastery levels. The probability of mastering attribute 11 tends to be .885. In comparison, attribute 1 tends to be an attribute relatively hard to master, while attribute 11 is relatively easy to master. Attribute 2 (i.e., Use ratio and rate reasoning to solve real-world and mathematical problems) is dependent on attribute 1, meaning that mastering attribute 1 positively contributes to mastering attribute 2. The results manifest that if a student has acquired attribute 1, this student may have a probability of .917 in mastering attribute 2, while if not, then the probability of mastering attribute 2 would drop to .085. Figure 17 shows the MCMC chain histories and posterior distributions for λ_1 and λ_2. The λ_1 distribution is nicely concentrated and centered at .493 with a small posterior variance, indicated by the path niggling around the estimated value and the density having thin tails and high peak. Similarly, the two levels of mastery status for attribute 2 also show paths concentrated and density centered tightly around the estimated values.

In another case, because attribute 9 has three prerequisite attributes (i.e., attributes 3, 4, 13), four levels of mastery status are possible: the probability of mastering attribute 9 with none of the prerequisite attributes mastered is .183, the mastery probability with one attribute mastered is .235, the mastery probability with two attributes mastered is .267, the mastery probability with all the prerequisite attributes mastered is .945. The four possibilities suggest that to understand and master attribute 9 (Apply and extend previous understandings of operations with factions to add, subtract, multiple, and divide rational numbers), it is better for students to first understand attribute 3 (Compute fluently with multi-digit numbers and find common factors and multiples), attribute 4 (Apply and extend previous understandings of numbers to the system of rational numbers), attribute 13 (Use equivalent faction as a strategy to add and subtract factions) all together. Lacking even one of the prerequisite attributes may largely reduce the probability of mastering attribute 9. Figure 18 shows the MCMC chain histories and posterior distributions for parameters associated with λ_9. It can be found that the MCMC chains for the first two levels (i.e., no or one prerequisite attributes being mastered) have relatively large variances around the estimated value, and their posterior *SD*s are relatively large, three to four times greater than the ones associated with the attribute. Further, for the other two levels of attribute 9 (i.e, two or all prerequisite attributes being mastered), although the MCMC chain is mixing well in the Bayesian analysis, they fail to quickly reach the target value after burn-ins, which leads to a larger posterior *SD* as well. Consistently, the density plots for these four levels show fatter tails on the first two levels and tighter ones for the latter two levels. These results may suggest that there is less evidence in this group to make extremely clear inferences on student mastery of attribute 9.

Item parameters. Table XXII also shows item parameter values of TIMSS data. Generally, each item has several parameters associated with different levels of mastery on the attributes measured by this item. Specifically, the items measuring only one attribute have two parameters: the probabilities of correctly answering this item when students fail to master the measured attribute (i.e., false positive) and when they have mastered the attribute (i.e., true positive). In other words, the false positive rate can be interpreted as the guessing parameter mostly used in CDMs, and similarly the probability deducted from 1 by the true positive rate can be interpreted as the slipping parameter. For example, as shown in Figure 20, the parameter for the false positive rate for item 14 measuring attribute 6 is mixing well in the MCMC trace plot and is tightly centered around .253, suggesting that students who fail to master attribute 6 would have a 25% probability of answering this item correctly; and the true positive rate is .721, suggesting that students who have mastered attribute 6 would have a 72% probability of answering this item correctly. In other words, the guessing parameter of this item is .251, and the slipping parameter is .279 (= 1 − .721). They are both relatively low, indicating a good quality of this item for identifying students who have and have not mastered attribute 6.

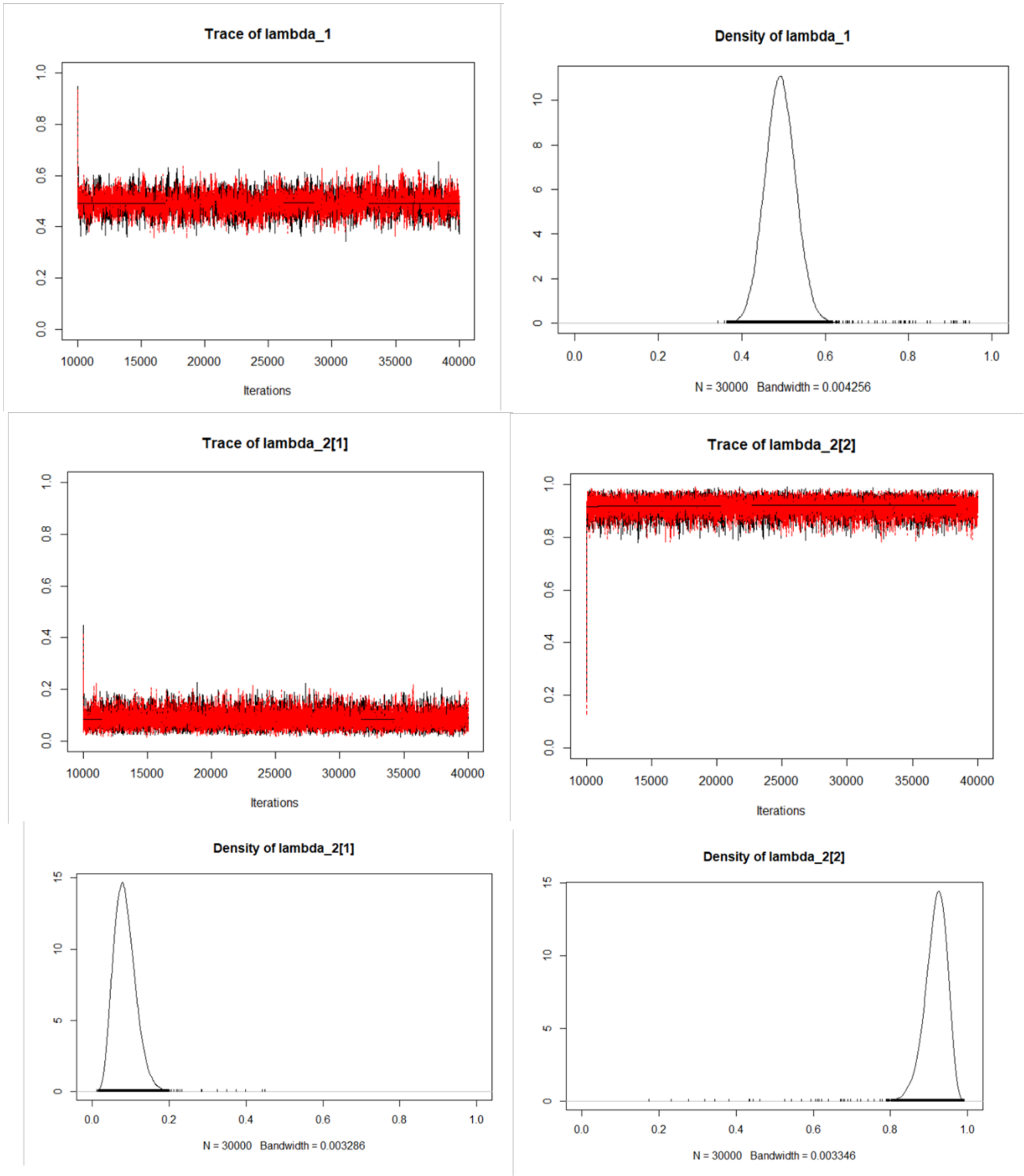


Figure 18. MCMC chain plots and density plots for parameters of λ_1, λ_2.

Note. The value in the bracket [a] denotes the *a*th parameter associated with the λ parameter.

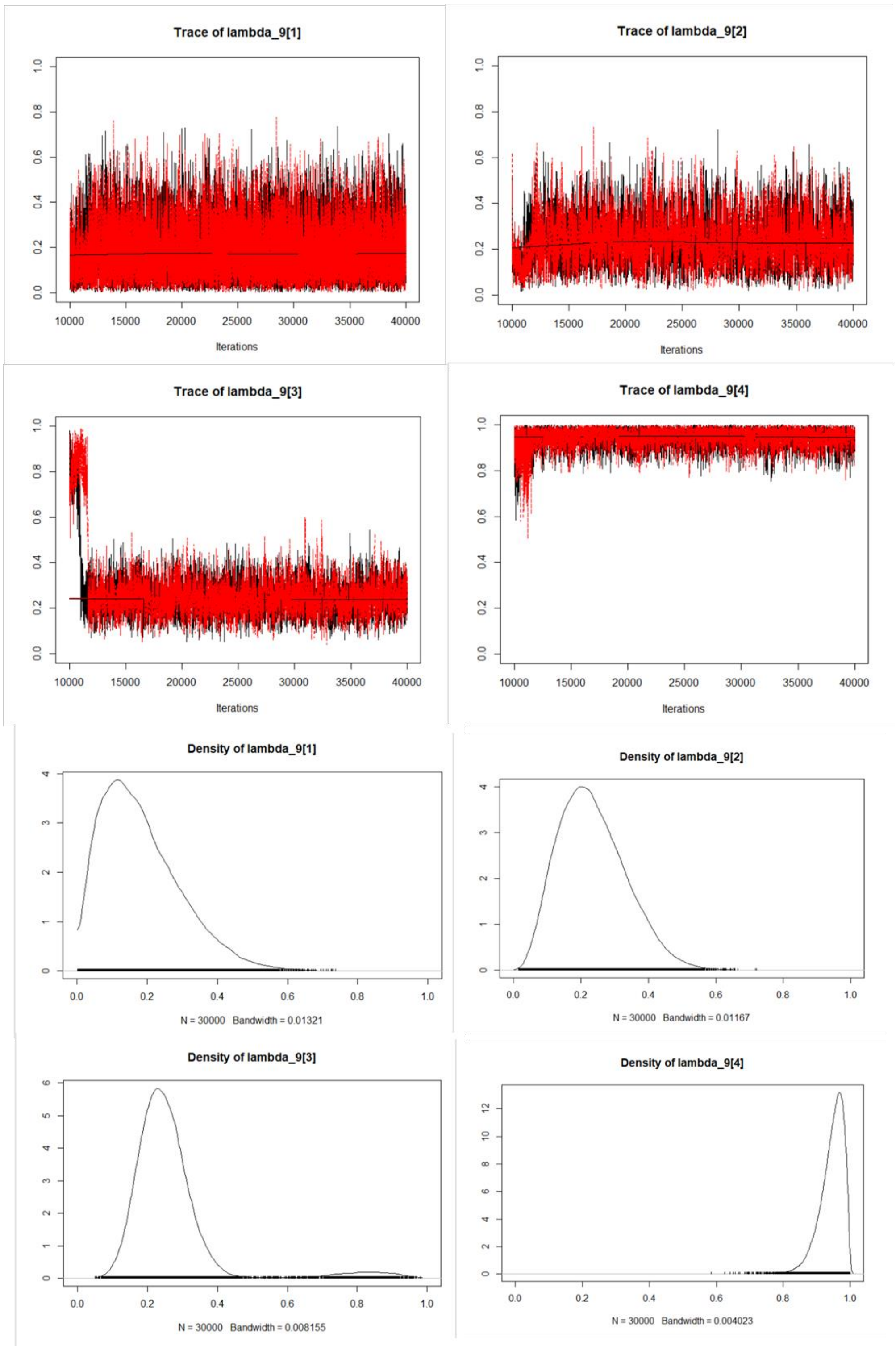


Figure 19. MCMC chain plots and density plots for parameters of λ_9.

Note. The value in the bracket [a] denotes the *a*th parameter associated with the λ parameter.

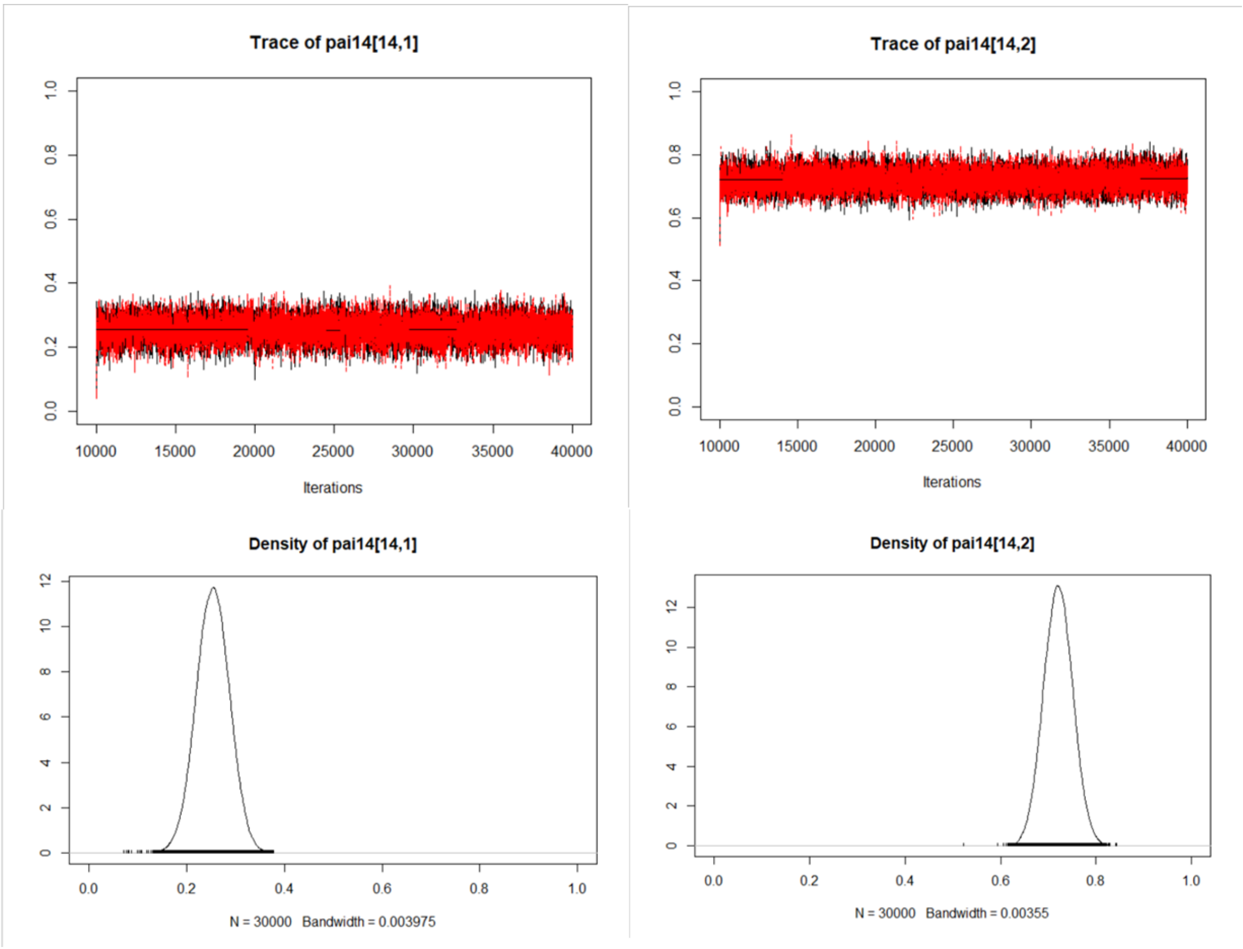


Figure 20. MCMC chain plots and density plots for parameters of π_{14}.

Note. The value in the bracket [b, a] denotes the *a*th parameter associated with the π_b parameter.

Figure 21 and Figure 22 show the MCMC chain histories and the density plots for the π parameters of Items 21 and 22, respectively. For both items, the MCMC chains of their π parameters are mixing well with small posterior *SD*s, and the posterior distributions are tightly concentrated around the parameter values. However, in terms of item quality, Item 21 has a guessing parameter of .154 and a slipping parameter of .646 (= 1 −.354), indicating that students who have mastered the measured attribute 5 may have a 64% probability of answering this item wrong. The high slipping parameter suggests that this item has low quality in identifying students who have mastered attribute 5. Item 22 has a guessing parameter of .492, suggesting that students who fail to master the measured attribute 2 have a 49.2% probability of answering this item correctly, and a slipping parameter of .093. It suggests that although this item has a low slipping parameter, its high guessing parameter marks this item as low-quality in differentiating students who fail to mastered attribute 5.

For another example, Item 8 measures attributes 5 and 9, therefore there are three mastery levels for this item: students who have mastered both attributes, students who have mastered one of the two attributes, and students who have not mastered any attributes. In this case, there is one parameter for each of the three levels. As shown in Figure 23 and Table XXII, the MCMC chains for these parameters are mixing well and their density plots are tightly centered around parameter values. Specifically, the probability of correctly answering item 8 for students who fail to master attributes 5 and 9 is .175; the correctness probability for students who have mastered either attribute 5 or 9 is .787; and as expected, the correctness probability for students who have mastered both attributes is .962. This result shows that mastering at least one attribute would have a relatively high probability of answering Item 8 correctly.

Table XXII lists the priors used for each parameter. Regarding the impact of the priors on the posterior distribution for both person and item parameters, it can be found from the MCMC chain histories and the discrepancy between priors and posterior estimates that the data are able to pull most of the BN posterior estimates away from the starting values. It suggests that the impact of priors on the BN posterior estimation is small for this analysis.

Item fit. As described in Section 3.3.5.1, the item fit indices for BN modeling are calculated by the PPMC, which can identify the discrepancy between the observed and the replicated item responses estimated by the model. In other words, it indicates what the data should look like if the model is true and how far their distribution is from the observed data distribution (Yan et al., 2003). Table XXIII summarizes the item-fit indices for a sample of 30000 iterations after 10000 burn-ins. According to the rule of thumb for the PPP value of PPMC, all the items fit well with no items having PPP values larger than .95 or smaller than .05. Most of them are around .5, indicating good fit.

Person fit. The person fit indices specifically explain whether students' item responses are consonant with their attribute mastery statuses. They allow us to know whether some responding behaviors fail to comply with the modeling estimation. Table XXIV presents the descriptive statistics of the person fit indices for 757 examinees in this data. As shown in Table XXIV, most of the person fit indices meet the criteria except that six students, .79% of the sample, have fit indices larger than .95 or smaller than .05, indicating that the person misfit of this model is negligible in estimating their mastery status of attributes.

I used six students as examples to unpack their performance in this assessment. Table XXV shows their item responses and their estimated mastery levels on each attribute. Among the six students, three have misfit indices and the others fit well by the model estimation. Table XXVI presents the mastery classifications based on the BN posterior probabilities of each attribute for these students.

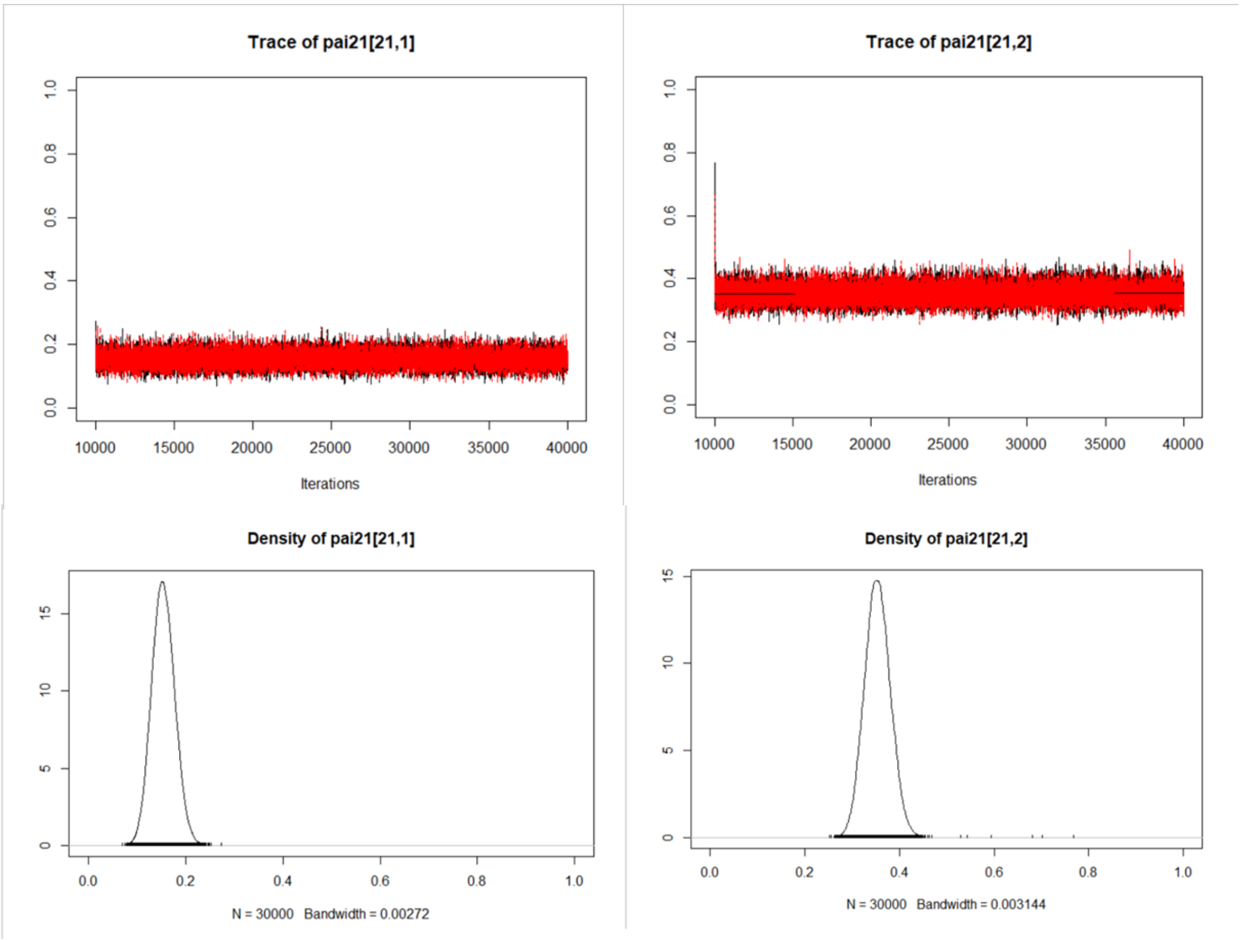


Figure 21. MCMC chain plots and density plots for parameters of π_{21}.

Note. The value in the bracket [b, a] denotes the *a*th parameter associated with the π_b parameter.

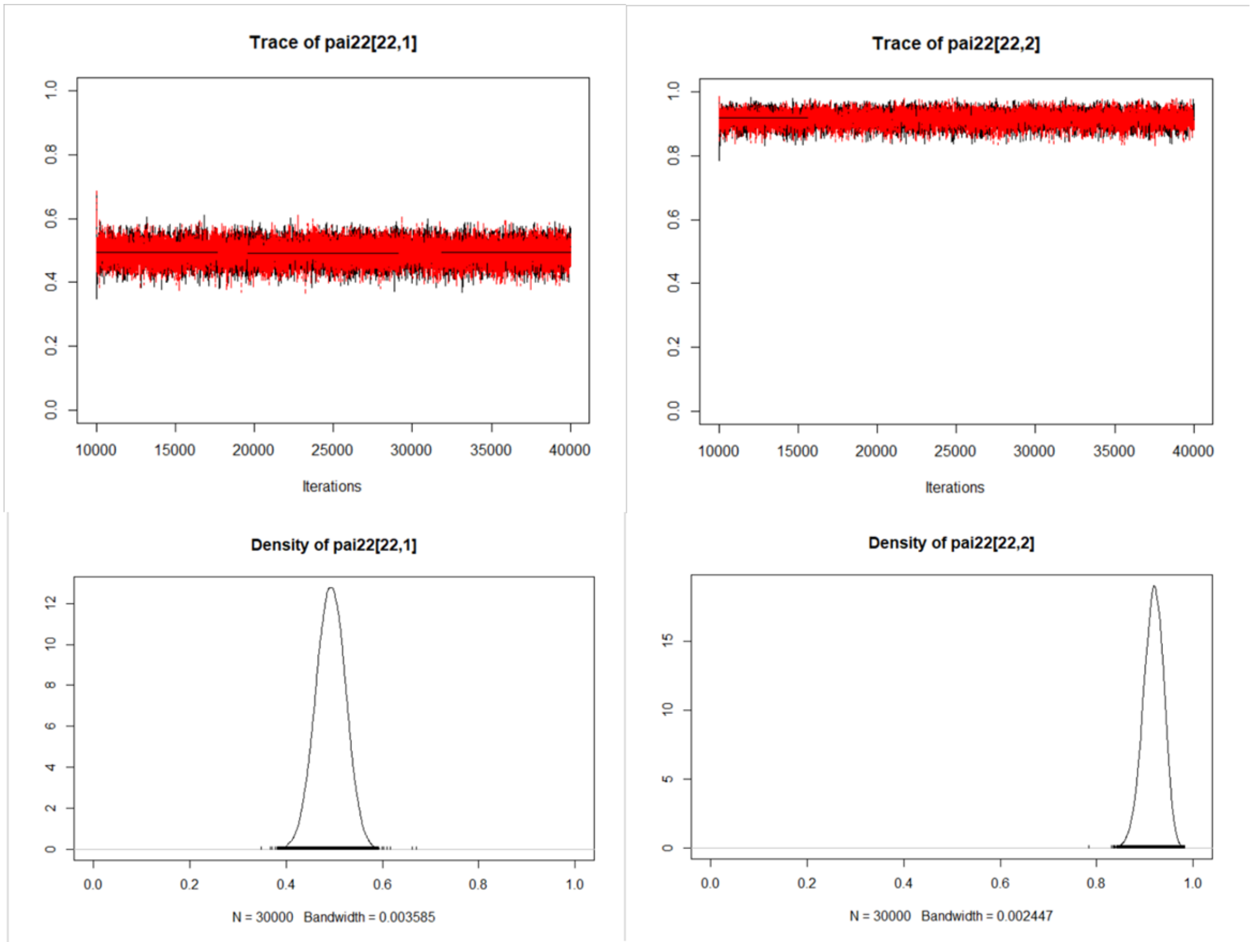


Figure 22. MCMC chain plots and density plots for parameters of π_{22}.

Note. The value in the bracket [b, a] denotes the *a*th parameter associated with the π_b parameter.

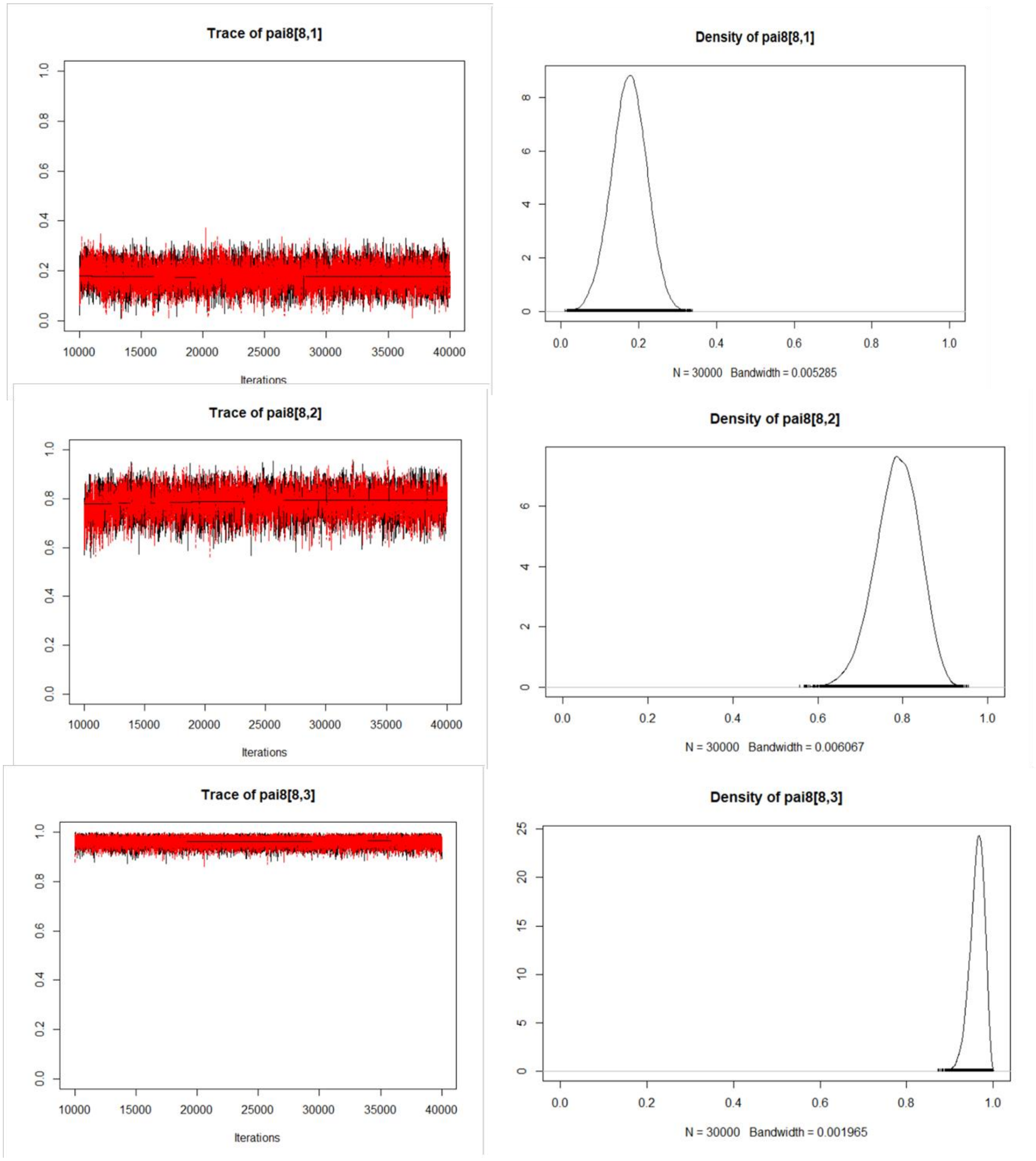


Figure 23. MCMC chain plots and density plots for parameters of π_8.

Note. The value in the bracket [b, a] denotes the *a*th parameter associated with the π_b parameter.

TABLE XXIII ITEM FIT INDICES FOR TIMSS DATA

Item	P(Obs >= Rep)
1	0.286
2	0.482
3	0.491
4	0.591
5	0.432
6	0.300
7	0.451
8	0.375
9	0.512
10	0.518
11	0.450
12	0.401
13	0.578
14	0.562
15	0.476
16	0.503
17	0.466
18	0.376
19	0.433
20	0.335
21	0.473
22	0.472
23	0.445

Note. The item fit criterion is that the *P* value is around .5 and is not larger than .95 or smaller than .05.

TABLE XXIV DESCRIPTIVE STATISTICS FOR PERSON FIT INDICES

Statistics	Value
Mean	0.492
Median	0.490
Standard Deviation	0.226
Minimum	0.037
Maximum	0.993

Note. The person fit criterion is that the *P* value is around .5 and is not larger than .95 or smaller than .05.

TABLE XXV EXAMINEE EXAMPLES OF ITEM RESPONSES

	Examinee						Q matrix: Attributes												
Item	23	208	481	279	331	213	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1	0	0	1	1	1	1										1		1
2	1	0	0	0	1	0						1							
3	1	0	0	1	0	0		1					1						
4	1	1	0	1	0	0				1					1				
5	1	0	0	0	0	0						1				1		1	
6	1	1	0	1	1	0						1	1						
7	1	0	0	1	1	0	1												
8	1	1	0	1	1	0					1				1				
9	1	0	1	1	1	1											1		
10	0	1	1	0	0	0						1							
11	1	1	1	1	1	0		1						1					
12	1	1	1	1	1	0			1										1
13	1	1	0	1	0	1					1								
14	1	1	0	1	1	0						1							
15	1	0	0	1	0	1		1										1	
16	1	1	0	1	0	0					1								
17	1	0	0	0	0	0				1									
18	1	1	0	0	1	0			1						1		1		1
19	0	1	1	1	0	0		1											
20	0	1	1	1	0	0	1												
21	0	1	0	1	1	0					1								
22	1	1	0	1	1	1		1											
23	1	1	0	1	0	0				1					1				

TABLE XXVI CLASSIFICATION RESULTS FOR EXAMINEE EXAMPLES

	Classification results for each attribute													
ID	1	2	3	4	5	6	7	8	9	10	11	12	13	Person fit
23	0	1	1	1	1	1	1	1	1	1	1	1	1	0.042
208	1	0	1	1	1	1	0	1	1	0	1	0	1	0.982
481	1	0	0	0	0	0	0	1	0	0	1	0	1	0.993
279	1	1	1	0	1	0	1	1	1	0	1	1	1	0.582
331	0	0	1	0	0	0	1	1	1	0	1	0	1	0.625
213	0	0	0	0	0	0	0	0	0	0	1	1	1	0.427

Examinee 23 shows a misfit pattern. It can be seen from their item responses that this student wrongly answered items 9, 19, 20, 21, which each measures attribute 6, 2, 1, 5, respectively. However, this student performed well on all other items including those measuring attributes 1, 2, 5, 6. These patterns are inconsistent for any of the attribute mastery status. Hence, this student has a less satisfactory person fit index and the results estimated from BN modeling, that this student has mastered all attribute except for attribute 1, tend to be less reliable.

Examinee 208 shows another misfit pattern that this student missed 8 out of 23 items. Specifically, this student missed items 2 and 17, which measure attribute 6 and 4 respectively. However, this student performed well on other harder items measuring attribute 6 alone (i.e., item 9 and 14 are all harder than item 2) and harder items measuring attribute 4 (i.e., item 4 and 23). In other words, this student missed relatively easy items while correctly answered hard items measuring the same attributes. This responding behavior is also inconsistent with a common attribute mastery status.

Similarly, Examinee 481 revealed misfit responding behavior. This student correctly answered the hard items 10, 19, 20 measuring attributes 1, 2, 6, respectively, but responded wrongly to the easy items measuring the same attributes (i.e., items 2, 7, 14, 22). This misfit pattern would make this estimation unreliable.

The rest three examinees are examples of having good person fit indices. Examinee 279 missed items 2, 5, 10, which measured attribute 6 alone or with other attributes, suggesting that this student had a lower probability of mastering attribute 6. According to the structural relationships, a lack of mastery on attribute 6 may contribute to a lower probability of mastering attribute 10. Consistently, this student answered Item 5 wrong which also measures attribute 10. Further, this student missed Item 17, which is the only item measuring attribute 4 alone, indicating a potential lack of mastery on attribute 4. These responding behaviors show few inconsistent patterns, and the corresponding estimated mastery status for this student, indicating a mastery of all attributes except for attributes 4, 6, 10, also showed the good fit of the attribute mastery status estimated by BN modeling.

Examinee 331 missed items 5, 10, 13, 15, 16, which measure attributes 12, 5, 6, 10 in some combination. These four attributes are also prerequisites to the next one in order according to the structural relationship. Hence, it is reasonable to find the estimated mastery status indicate a lack of mastery on these four attributes for this student. Further, this student wrongly answered items 19, 20, 23 measuring attributes 1, 2, 4, respectively, suggesting a possible lack of these attributes. These patterns are consistent with the estimated attribute classification by BN modeling as well.

Finally, Examinee 213 tended to master attributes 11, 12, 13 as this student correctly answered items 1, 9, 15, 22. Although the correct answers to items 15 and 22 may also indicate a potential mastery of attribute 2, this student failed to answer all other questions measuring attribute 2. Hence, it is not surprising to find that the estimated mastery status of this student excludes attribute 2.

I further conducted a correlation analysis between the raw score of each student and the number of mastered attributes in their mastery profile estimated by BN. The correlation estimate of .90 ($p < .001$) suggests a positively and highly consistent relationship between the raw score and the estimated mastery of attributes. This finding indicates that students who answered more items correctly would be estimated to have mastered more attributes by BN. It further manifests the capacity of BN modeling in adequately classifying students into mastery levels of each attribute.

To display how attributes are related to each other and how item responses are related to student attribute mastery status, I used Netica to draw the graphical display to showcase one part of these relationships. To make a straightforward and clear graph, I used attributes 5, 6, 10, 12 and the items measuring these attributes as an illustration. As shown in Figure 24, around half of students tend to master attributes 5, 6, and 10, while 70% may have mastered attribute 12. With respect to item difficulties, students may have a 50% probability of providing correct responses to items 5, 13, 14, 16, and students seem to have only around a 30% probability of giving correct responses to items 10, 21, suggesting these two items are relatively difficult. Item 2 tends to be an easy item with a correctness probability of around 70%.

Figure 25 shows the mastery profile estimated by BN and item responses of Examinee 331. It can be found that this student may have a 50% probability of mastering attribute 12, while this student tends to have a relatively low likelihood to master attributes 5, 6, and 10. Viewing the relationships from another perspective, if we assume a student has mastered attributes 5 and 12 but not attributes 6 and 10 as shown in Figure 25, this student's item responses can also be predicted by the posterior distribution among attributes and items. Specifically, this student would have a relatively low likelihood of correctly answering items 2, 5, 10, 14, and 21, while this student would have a high probability of correctly answering items 13 and 16. Generally, the graphical display of the structure in attributes and its connection with items would help students to understand their current mastery levels and customize specific learning plans on how they should achieve improvement and what components they should focus on in remediation. It would also contribute to teachers' instructional design from both individualized and class levels on how to improve students' performance.

To achieve the ultimate goal of a diagnostic assessment, BN can be used to provide a score report with formative feedback suggesting individualized learning paths for students and potential instructional strategies for teachers. I used the three students with good person fit indices to showcase diagnostic score reports. Figure 26 shows a diagnostic report for examinee 279. The report lists the attributes this student may not have mastered and the items they missed to provide students with a checklist to review. Based on the structural relationships among attributes, I also made some suggestions on future learning paths or instructional strategies that may help students and teachers to prioritize their tasks. Figure 27 shows the diagnostic report for examinee 311. This report involves more suggestions on the learning paths for attributes with structural relationships as this student failed to master more attributes.

In addition to the individual diagnostic report, it would be practical to provide teachers with a group-level diagnostic report. Figure 28 displays the distribution of mastery levels for each attribute and the structure in attributes for the entire sample, which teachers can use to adjust their teaching. The diagrams in Figure 28 indicate students' mastery of each attribute and suggest sequences of attribute teaching. Specifically, the hierarchy of attributes 12-5-6-10 follows the assumption that the prerequisite (top) attributes show a higher mastery probability than the advanced (bottom) attributes. Similarly, the structure among attributes 13-3-11-9-4 mostly follows the hypothesis that attributes 13, 3 and 11 as the basic attributes tend to be easier for students to master than attributes 9 and 4. Although attribute 9 is dependent on attribute 4, it seems easier to master than attribute 4. It also occurred among the hierarchy of attributes 1-2-7-8 that, inconsistent with the hypothesis, the prerequisite attributes 1 and 2 seem to be difficult for students to master than its descendant attributes. These inconsistencies occurred might be due to two reasons: 1) the structure in attributes is inaccurate so that the posterior distribution among attributes fails to present the hypothesized structure; 2) if the definitions of attributes demonstrate that the structure should be correct, the inconsistencies occurred probably because the items are not well designed to measure the target attributes or the number of items are insufficient to fully tap students' performance on the measured attributes. In the TIMSS data, the definitions of attributes 1 and 2 tend to be the prerequisite attributes for attributes 7 and 8. In this case, following the fact that only two items measuring attribute 7 or 8 together with other attributes, the sparse Q matrix might cause that this assessment fails to measure attributes 7 and 8 using enough items.

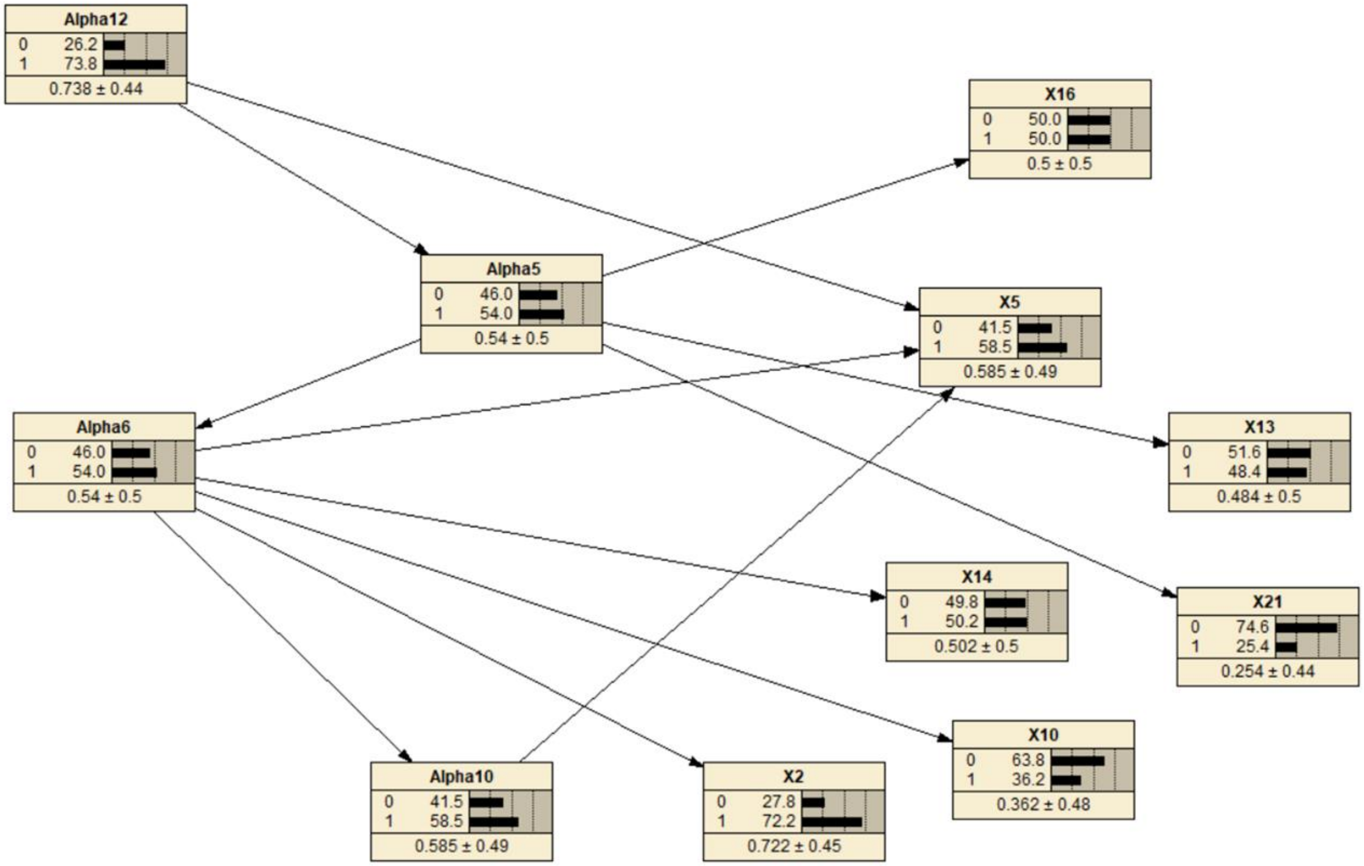


Figure 24. Graphical display of the BN posterior distribution among five selected attributes and the items measuring them.

Note. Alpha denotes the mastery classification for each attribute, *X* denotes the items responses to each item.

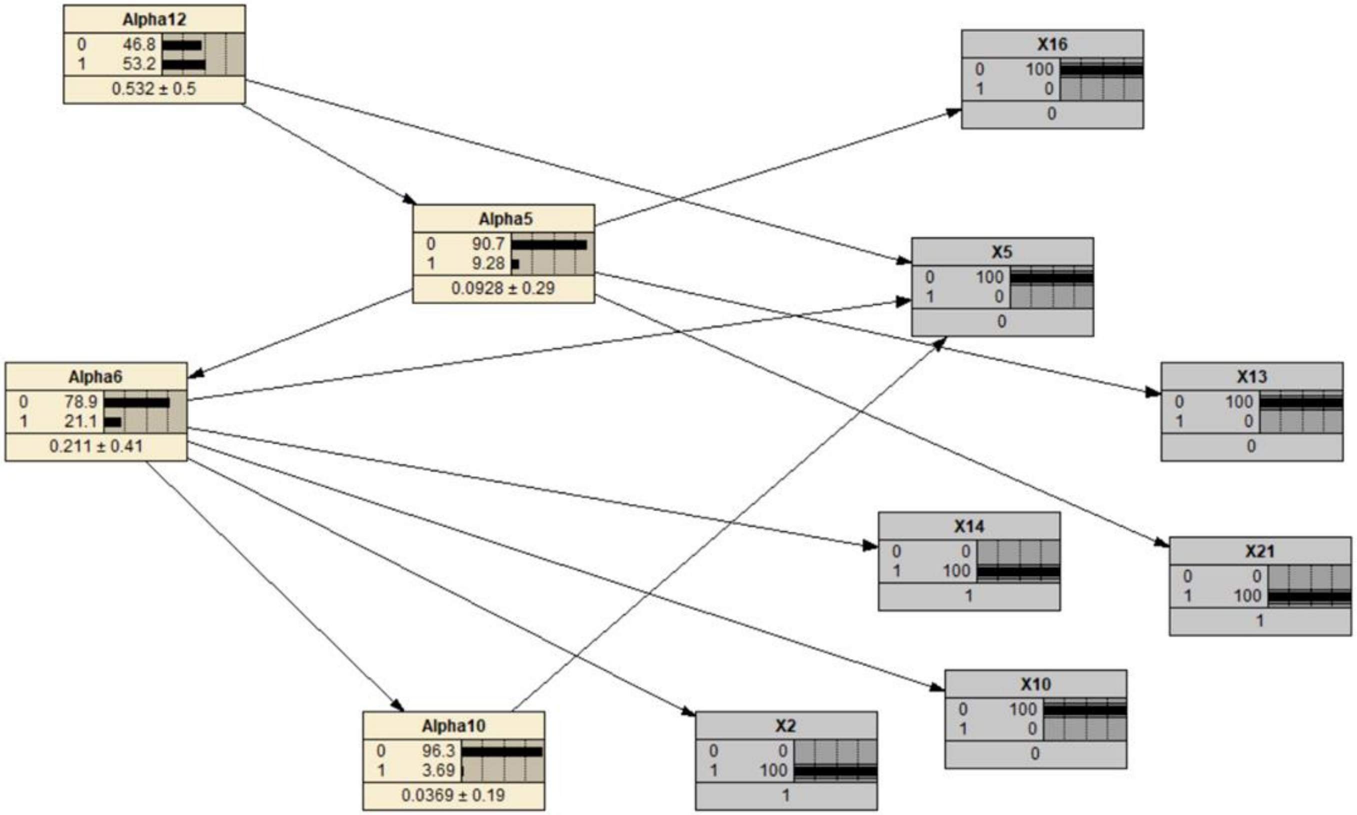
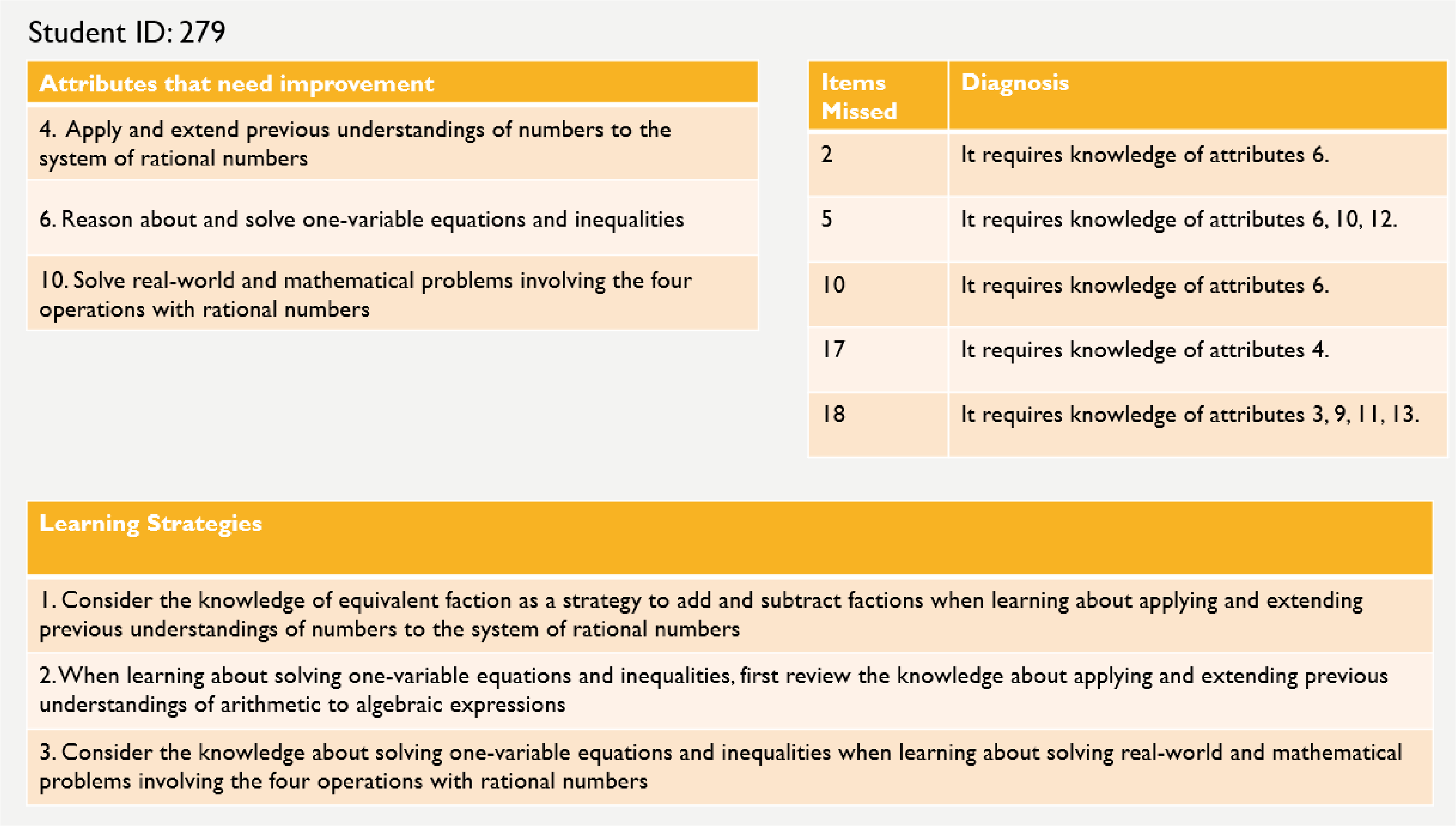


Figure 25. Graphical display of the mastery status of five selected attributes for Examine 331.

Note. Alpha denotes the mastery classification for each attribute, *X* denotes the items responses to each item.



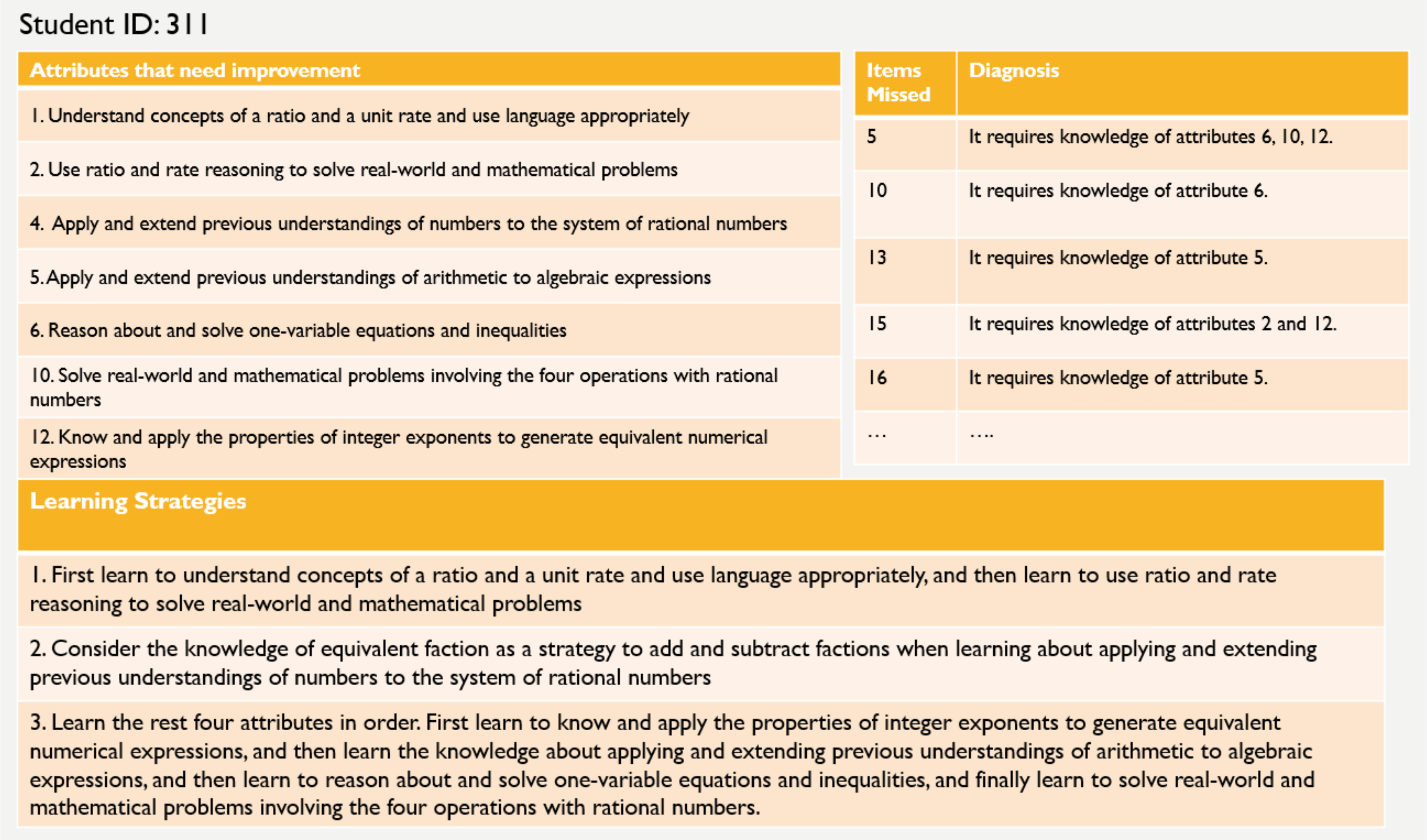
Student ID: 279

Attributes that need improvement
4. Apply and extend previous understandings of numbers to the system of rational numbers
6. Reason about and solve one-variable equations and inequalities
10. Solve real-world and mathematical problems involving the four operations with rational numbers

Items Missed	Diagnosis
2	It requires knowledge of attributes 6.
5	It requires knowledge of attributes 6, 10, 12.
10	It requires knowledge of attributes 6.
17	It requires knowledge of attributes 4.
18	It requires knowledge of attributes 3, 9, 11, 13.

Learning Strategies
1. Consider the knowledge of equivalent faction as a strategy to add and subtract factions when learning about applying and extending previous understandings of numbers to the system of rational numbers
2. When learning about solving one-variable equations and inequalities, first review the knowledge about applying and extending previous understandings of arithmetic to algebraic expressions
3. Consider the knowledge about solving one-variable equations and inequalities when learning about solving real-world and mathematical problems involving the four operations with rational numbers

Figure 26. The diagnostic report for Examinee 279.



Student ID: 311

Attributes that need improvement
1. Understand concepts of a ratio and a unit rate and use language appropriately
2. Use ratio and rate reasoning to solve real-world and mathematical problems
4. Apply and extend previous understandings of numbers to the system of rational numbers
5.Apply and extend previous understandings of arithmetic to algebraic expressions
6. Reason about and solve one-variable equations and inequalities
10. Solve real-world and mathematical problems involving the four operations with rational numbers
12. Know and apply the properties of integer exponents to generate equivalent numerical expressions

Items Missed	Diagnosis
5	It requires knowledge of attributes 6, 10, 12.
10	It requires knowledge of attribute 6.
13	It requires knowledge of attribute 5.
15	It requires knowledge of attributes 2 and 12.
16	It requires knowledge of attribute 5.
…	….

Learning Strategies

1. First learn to understand concepts of a ratio and a unit rate and use language appropriately, and then learn to use ratio and rate reasoning to solve real-world and mathematical problems

2. Consider the knowledge of equivalent faction as a strategy to add and subtract factions when learning about applying and extending previous understandings of numbers to the system of rational numbers

3. Learn the rest four attributes in order. First learn to know and apply the properties of integer exponents to generate equivalent numerical expressions, and then learn the knowledge about applying and extending previous understandings of arithmetic to algebraic expressions, and then learn to reason about and solve one-variable equations and inequalities, and finally learn to solve real-world and mathematical problems involving the four operations with rational numbers.

Figure 27. The diagnostic report for Examinee 311.

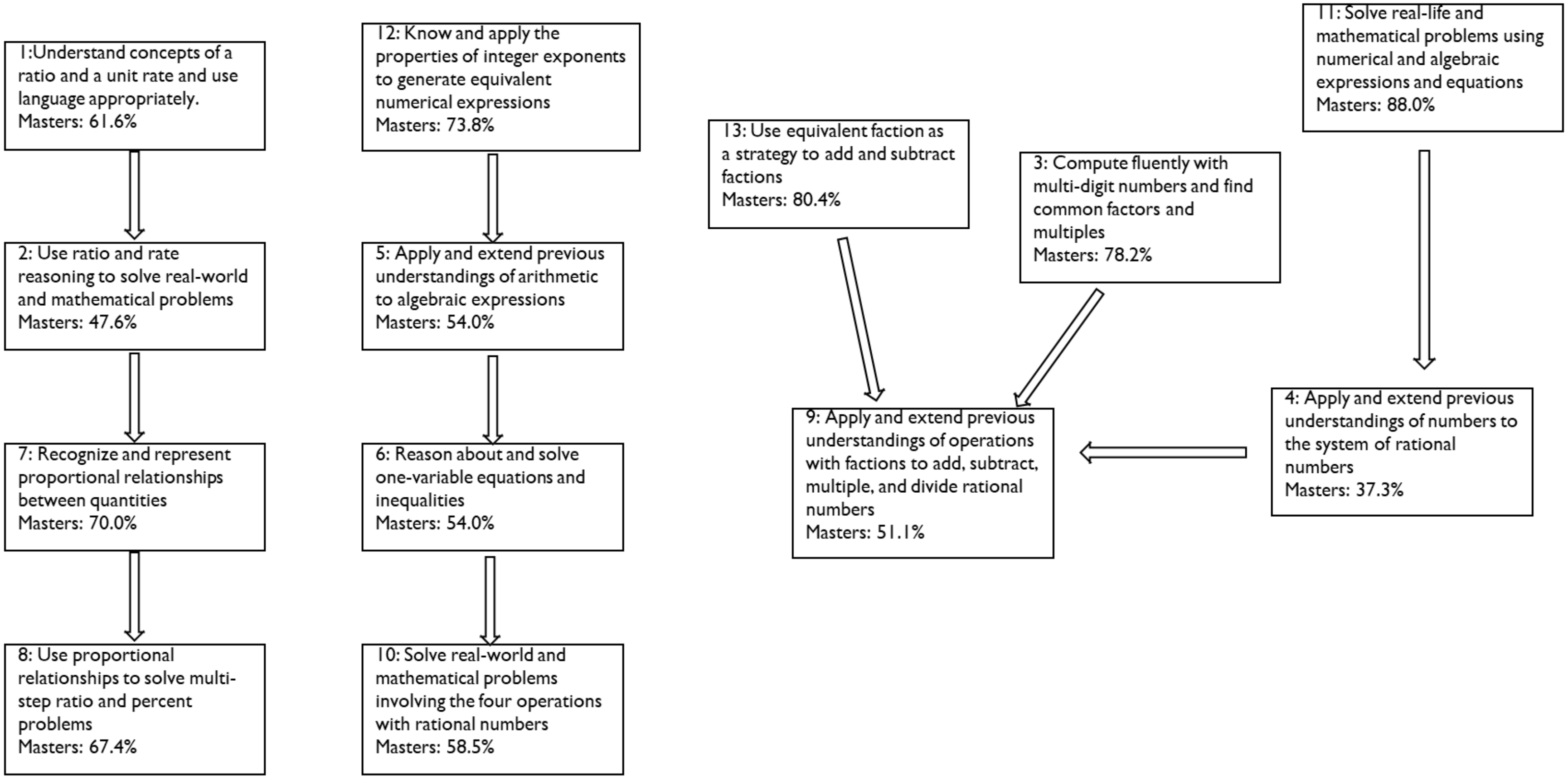

1:Understand concepts of a ratio and a unit rate and use language appropriately.
Masters: 61.6%
2: Use ratio and rate reasoning to solve real-world and mathematical problems
Masters: 47.6%
7: Recognize and represent proportional relationships between quantities
Masters: 70.0%
8: Use proportional relationships to solve multi-step ratio and percent problems
Masters: 67.4%
12: Know and apply the properties of integer exponents to generate equivalent numerical expressions
Masters: 73.8%
5: Apply and extend previous understandings of arithmetic to algebraic expressions
Masters: 54.0%
6: Reason about and solve one-variable equations and inequalities
Masters: 54.0%
10: Solve real-world and mathematical problems involving the four operations with rational numbers
Masters: 58.5%
13: Use equivalent faction as a strategy to add and subtract factions
Masters: 80.4%
3: Compute fluently with multi-digit numbers and find common factors and multiples
Masters: 78.2%
11: Solve real-life and mathematical problems using numerical and algebraic expressions and equations
Masters: 88.0%
9: Apply and extend previous understandings of operations with factions to add, subtract, multiple, and divide rational numbers
Masters: 51.1%
4: Apply and extend previous understandings of numbers to the system of rational numbers
Masters: 37.3%

Figure 28. The group-level report of each attribute for the TIMSS data.

Note. Percentage refers to the percent of students who have mastered each of the attributes

4.2.3 <u>Polytomous Attribute Data Analysis</u>

As no real data is available for tests with polytomous attributes, I used the polytomous attribute data for the polytomous GDINA Model simulated by Chen and de la Torre (2013). The data have 1000 students for a test of 30 dichotomously-score items and five polytomous attributes. The Q matrix is shown in Table XXVII.

Priors. The starting values for the polytomous attribute data are slightly different from the TIMSS data, especially for the person parameters. In the case of polytomous attributes with three levels of non-mastery denoted by 0, medium mastery denoted by 1, high mastery denoted by 2, the person parameters λ would follow a Dirichlet distribution, which would provide a probability for each mastery level of attributes. Also, as the polytomous attributes in this dataset are not assumed to have a structural relationship, attributes are considered independently contributing to item responses. Based on all the information, the priors for the probabilities of the three mastery levels for all attributes follow Dirichlet (2,6,2), which are .200, .606, and .193.

Regarding priors for item parameters, I classified them into three evidence model groups based on the number of attributes they measure. First, for the items measuring only one attribute, the probabilities of correctness for students who have and have not reached the required mastery level of the measured attribute are assumed to follow Beta (8,2) and Beta (2,8), respectively. Items measuring two attributes were hypothesized to have probabilities of correctness for students who have reached the mastery levels of none, one, two measured attributes following Beta (2,8), Beta (5,5), Beta (8,2), respectively. Items measuring three attributes were assigned to follow Beta (2,8), Beta (4,6), Beta (6,4), Beta (8,2) for their probabilities of correctness for student who have reached the mastery levels of none, one, two, and three measured attributes, respectively.

TABLE XXVII Q MATRIX FOR THE TEST DATA WITH POLYTOMOUS ATTRIBUTES

Item Number	A1	A2	A3	A4	A5
1	1	0	0	0	0
2	0	1	0	0	0
3	0	0	1	0	0
4	0	0	0	1	0
5	0	0	0	0	1
6	1	2	0	0	0
7	0	1	2	0	0
8	0	0	1	2	0
9	0	0	0	1	1
10	2	0	0	0	1
11	2	2	0	1	0
12	2	1	0	0	1
13	1	0	2	2	0
14	0	2	1	0	1
15	0	0	2	2	1
16	2	0	0	0	0
17	0	2	0	0	0
18	0	0	2	0	0
19	0	0	0	2	0
20	0	0	0	0	1
21	2	0	2	0	0
22	0	2	0	2	0
23	0	0	2	0	1
24	2	0	0	2	0
25	0	2	0	0	1
26	1	0	0	1	1
27	0	1	1	1	0
28	1	1	1	0	0
29	0	1	0	1	1
30	1	0	1	0	1

Note. This Q matrix was described in Chen and de la Torre (2013).

Parameter values. Table XXVIII summarizes the statistics of the estimated person and item parameters, their interquartile range, their $\hat{R}$ statistics resulting from 6000 iterations following 1500 burn-ins, and priors. As can be seen, all parameters have $\hat{R}$ statistics smaller than 1.2, suggesting the estimation of these parameters has converged.

Person parameters. In the polytomous-attribute case, the interpretation of each attribute mastery level is different from the dichotomous case. In this data, students are classified into non-mastery, medium mastery, high mastery on each attribute. As can be seen Table XXVIII three parameters are associated with each attribute, and each specifies the probability of mastering the attribute in the non-mastery, medium mastery, high mastery levels, respectively. I used attributes 3 and 5 for an illustration to explain their parameter values and MCMC chain mixing. We may find that the non-mastery probability for attribute 3 is .167, the medium mastery probability is .585, the high mastery probability is .248. As shown in Figure 29, the MCMC chains for the three parameters are mixing well with small variance and niggling around the parameter values. The density plots also indicate the posterior distribution of each parameter is tightly concentrated around the estimated value.

For another example, the mastery probabilities of attribute 5 reveal that attribute 5 is relatively harder to master compared to attribute 3. Specifically, the non-mastery probability is around .492, the medium mastery probability is .372, and the high mastery probability is .136. As indicated in Figure 30, the probability of non-mastery is mixing well with a tightly concentrated posterior distribution. However, the probabilities of medium and high mastery are not mixing very well with large posterior *SD*s which leads to the fluctuation of MCMC chains in a relatively large range. Consistently, the density plots have fatter tails. It may be because the number of students mastering attribute 5 in either medium or high level is small compared to other attributes, hence less evidence was provided for BN modeling to yield more stable parameter values.

TABLE XXVIII PARAMETER VALUES FOR THE POLYTOMOUS ATTRIBUTE DATA

Parameters	Mean	SD	25.00%	Median	75.00%	Rhat	Priors
lambda_1[1]	0.288	0.016	0.277	0.287	0.299	1.001	0.200
lambda_1[2]	0.510	0.019	0.497	0.510	0.522	1.001	0.606
lambda_1[3]	0.203	0.015	0.192	0.203	0.213	1.001	0.193
lambda_2[1]	0.251	0.019	0.239	0.251	0.263	1.001	0.200
lambda_2[2]	0.501	0.023	0.485	0.501	0.516	1.003	0.606
lambda_2[3]	0.248	0.019	0.235	0.248	0.261	1.002	0.193
lambda_3[1]	0.167	0.020	0.153	0.166	0.180	1.002	0.200
lambda_3[2]	0.585	0.026	0.568	0.586	0.603	1.003	0.606
lambda_3[3]	0.248	0.020	0.235	0.247	0.261	1.001	0.193
lambda_4[1]	0.262	0.017	0.250	0.262	0.273	1.001	0.200
lambda_4[2]	0.547	0.021	0.533	0.546	0.561	1.001	0.606
lambda_4[3]	0.191	0.016	0.180	0.191	0.202	1.001	0.193
lambda_5[1]	0.492	0.018	0.481	0.492	0.503	1.008	0.200
lambda_5[2]	0.372	0.077	0.321	0.382	0.431	1.001	0.606
lambda_5[3]	0.136	0.076	0.076	0.125	0.186	1.001	0.193
pai[1,1]	0.067	0.026	0.050	0.065	0.083	1.001	0.196
pai[2,1]	0.024	0.008	0.019	0.024	0.029	1.001	0.196
pai[3,1]	0.048	0.022	0.033	0.045	0.061	1.001	0.196
pai[4,1]	0.261	0.018	0.249	0.261	0.273	1.002	0.196
pai[5,1]	0.125	0.053	0.087	0.123	0.160	1.002	0.196
pai[6,1]	0.033	0.010	0.026	0.032	0.039	1.001	0.196
pai[7,1]	0.024	0.019	0.012	0.021	0.032	1.001	0.196
pai[8,1]	0.012	0.006	0.007	0.011	0.015	1.001	0.196
pai[9,1]	0.104	0.015	0.093	0.103	0.113	1.003	0.196
pai[10,1]	0.156	0.018	0.144	0.156	0.168	1.001	0.196
pai[1,2]	0.989	0.009	0.986	0.991	0.994	1.010	0.800
pai[2,2]	0.894	0.039	0.869	0.897	0.922	1.001	0.800
pai[3,2]	0.816	0.019	0.803	0.816	0.828	1.001	0.800
pai[4,2]	0.769	0.036	0.745	0.769	0.793	1.002	0.800
pai[5,2]	0.841	0.017	0.830	0.841	0.852	1.002	0.800
pai[6,2]	0.748	0.043	0.719	0.748	0.777	1.002	0.800
pai[7,2]	0.887	0.017	0.877	0.888	0.898	1.002	0.800
pai[8,2]	0.784	0.045	0.755	0.785	0.815	1.001	0.800
pai[9,2]	0.974	0.014	0.968	0.975	0.981	1.001	0.800
pai[10,2]	0.972	0.013	0.967	0.973	0.978	1.003	0.800
pai[11,1]	0.206	0.027	0.187	0.205	0.224	1.001	0.196
pai[12,1]	0.161	0.015	0.150	0.160	0.171	1.001	0.196
pai[13,1]	0.231	0.034	0.207	0.230	0.253	1.007	0.196
pai[14,1]	0.145	0.014	0.136	0.145	0.155	1.001	0.196
pai[15,1]	0.348	0.021	0.334	0.348	0.361	1.002	0.196
pai[16,1]	0.063	0.027	0.044	0.061	0.080	1.002	0.196
pai[17,1]	0.376	0.024	0.359	0.375	0.392	1.001	0.196
pai[18,1]	0.082	0.015	0.071	0.081	0.092	1.001	0.196
pai[19,1]	0.203	0.021	0.189	0.202	0.217	1.001	0.196
pai[20,1]	0.087	0.022	0.072	0.086	0.101	1.001	0.196
pai[11,2]	0.223	0.023	0.207	0.222	0.238	1.001	0.503
pai[12,2]	0.164	0.028	0.145	0.163	0.181	1.001	0.503
pai[13,2]	0.275	0.023	0.259	0.275	0.290	1.001	0.503
pai[14,2]	0.151	0.026	0.133	0.150	0.167	1.001	0.503
pai[15,2]	0.363	0.032	0.342	0.363	0.384	1.002	0.503

pai[16,2]	0.082	0.014	0.072	0.081	0.091	1.002	0.503
pai[17,2]	0.332	0.023	0.317	0.332	0.347	1.001	0.503
pai[18,2]	0.090	0.015	0.079	0.089	0.099	1.002	0.503
pai[19,2]	0.171	0.021	0.157	0.171	0.185	1.001	0.503
pai[20,2]	0.128	0.021	0.114	0.127	0.141	1.001	0.503
pai[11,3]	0.946	0.022	0.933	0.948	0.962	1.001	0.800
pai[12,3]	0.891	0.039	0.866	0.893	0.919	1.001	0.800
pai[13,3]	0.941	0.025	0.925	0.942	0.959	1.001	0.800
pai[14,3]	0.880	0.042	0.854	0.885	0.910	1.002	0.800
pai[15,3]	0.838	0.051	0.805	0.841	0.875	1.001	0.800
pai[16,3]	0.854	0.034	0.832	0.855	0.878	1.001	0.800
pai[17,3]	0.920	0.026	0.904	0.922	0.938	1.001	0.800
pai[18,3]	0.775	0.043	0.746	0.776	0.805	1.002	0.800
pai[19,3]	0.837	0.035	0.814	0.839	0.862	1.001	0.800
pai[20,3]	0.855	0.019	0.844	0.855	0.867	1.001	0.800
pai[21,1]	0.056	0.038	0.028	0.048	0.075	1.001	0.196
pai[22,1]	0.156	0.030	0.136	0.155	0.176	1.001	0.196
pai[23,1]	0.212	0.103	0.135	0.201	0.280	1.001	0.196
pai[24,1]	0.260	0.031	0.239	0.259	0.281	1.001	0.196
pai[25,1]	0.202	0.040	0.175	0.200	0.228	1.001	0.196
pai[26,1]	0.357	0.083	0.302	0.356	0.411	1.001	0.196
pai[27,1]	0.052	0.028	0.032	0.048	0.068	1.001	0.196
pai[28,1]	0.234	0.049	0.201	0.232	0.266	1.001	0.196
pai[29,1]	0.299	0.067	0.254	0.296	0.343	1.002	0.196
pai[30,1]	0.308	0.026	0.291	0.308	0.326	1.002	0.196
pai[21,2]	0.087	0.027	0.068	0.085	0.104	1.001	0.404
pai[22,2]	0.165	0.018	0.152	0.165	0.177	1.003	0.404
pai[23,2]	0.182	0.046	0.151	0.181	0.212	1.003	0.404
pai[24,2]	0.245	0.022	0.230	0.244	0.260	1.001	0.404
pai[25,2]	0.199	0.024	0.183	0.198	0.215	1.001	0.404
pai[26,2]	0.327	0.041	0.299	0.326	0.354	1.002	0.404
pai[27,2]	0.053	0.015	0.043	0.052	0.061	1.001	0.404
pai[28,2]	0.171	0.029	0.151	0.170	0.189	1.003	0.404
pai[29,2]	0.136	0.027	0.117	0.134	0.153	1.001	0.404
pai[30,2]	0.341	0.029	0.322	0.340	0.360	1.001	0.404
pai[21,3]	0.102	0.028	0.084	0.100	0.118	1.001	0.594
pai[22,3]	0.196	0.035	0.173	0.195	0.218	1.001	0.594
pai[23,3]	0.151	0.030	0.132	0.150	0.168	1.001	0.594
pai[24,3]	0.238	0.031	0.217	0.237	0.258	1.002	0.594
pai[25,3]	0.204	0.023	0.189	0.204	0.219	1.001	0.594
pai[26,3]	0.362	0.026	0.344	0.361	0.380	1.001	0.594
pai[27,3]	0.057	0.013	0.047	0.056	0.065	1.001	0.594
pai[28,3]	0.174	0.024	0.158	0.174	0.189	1.001	0.594
pai[29,3]	0.193	0.027	0.175	0.192	0.210	1.001	0.594
pai[30,3]	0.390	0.037	0.365	0.389	0.414	1.001	0.594
pai[21,4]	0.845	0.020	0.833	0.846	0.858	1.001	0.800
pai[22,4]	0.922	0.032	0.902	0.925	0.945	1.001	0.800
pai[23,4]	0.786	0.023	0.771	0.786	0.800	1.005	0.800
pai[24,4]	0.935	0.034	0.916	0.940	0.960	1.001	0.800
pai[25,4]	0.782	0.039	0.757	0.784	0.810	1.001	0.800
pai[26,4]	0.809	0.023	0.795	0.810	0.824	1.001	0.800
pai[27,4]	0.696	0.045	0.666	0.696	0.727	1.002	0.800
pai[28,4]	0.977	0.014	0.972	0.978	0.983	1.001	0.800

pai[29,4]	0.932	0.018	0.924	0.934	0.943	1.001	0.800
pai[30,4]	0.922	0.036	0.901	0.927	0.949	1.001	0.800

Note. lambda [a] denotes the a_{th} mastery level of the λ parameter, where 1 denotes non-mastery, 2 denotes medium mastery, 3 denotes high mastery. Pai[b,a] denotes the a_{th} parameter associated with Item b, where 1 denotes none of the required mastery levels in the measured attributes are reached, 2 denotes one of the required mastery levels in the measured attributes is reached, etc.

Item parameters. Different from the dichotomous case, the item parameters in the polytomous attribute case would evaluate not only whether students have mastered the measured attributes but also how well they master each attribute. I used three item examples to explain their item parameter values. Figure 31 shows that all three parameters of Item 17, which measures two attributes, are mixing well with tight posterior distributions. We may find that students who have failed to reach the required mastery levels of the two attributes or have only reached the required mastery level for one attribute would have the similar correctness probabilities around .3, while the probability of answering Item 17 correctly is .920 once students have reached the required mastery level for all the measured attributes. These results suggest that Item 17 may have a relatively high guessing parameter, as students who fail to reach the required mastery levels of both attributes still have 30% probability of answer this item correctly.

For Item 18, students who have failed to reach the required mastery levels on none or one attribute would have a very low probability of correctly answering this item. Students who have reached the measured mastery level of both measured attributes would have a .775 probability of answering Item 18 correctly. It reveals that Item 18 is harder than Item 17. The MCMC chains for all the three parameters are mixing well with tightly concentrated posterior distributions.

Finally, Item 30 measures three attributes so that it has four parameters: the probabilities of giving a correct response for students who have achieved the measured mastery levels of none, one, two, or three attributes. As can be seen in Figure 32, all the parameter values are mixing well, and the posterior distributions are tightly centered around the estimated parameter values.

Item fit and person fit. Table XXIX specifies the item fit indices, indicating that all the items fit well. Table XXX shows the descriptive statistics of the person fit indices for 1000 examinees, which indicates that around 4% = (40/1000) of students have misfit indices. The overall model fit is .849.

Attribute mastery classification. The correlation between students' raw score and students' sum of mastery levels on each attribute is high, r = .915 (p < .01). I then used a student example to demonstrate the diagnostic report in a polytomous attribute case (see in Figure 33).

Table XXXI specifies the group-level distribution for this polytomous data. It can be noted that the proportion of median mastery for all attributes is around 50% in this group. The proportions of non-mastery and high mastery would suggest attribute difficulties. For example, attributes 1, 2, 5 tend to have higher probabilities of non-mastery and lower likelihood of high mastery, in which attribute 5 is the most difficult attribute and that of high mastery is almost zero. This type of group information would help teachers to learn about the difficulty of each attribute for the class and to design and adjust their instructional strategies to cater to their student's learning needs.

Taken together, I used the BN approach to analyze two existing datasets. The results illustrate the good model fitting of the BN approach and its capacity of producing model parameters with good fit. Based on the reliable model parameters, the graphical display of the joint distribution over latent variables and item responses would contribute to diagnostic reports of person-level and group-level student performance for teaching and learning purposes.

TABLE XXIX ITEM FIT INDICES FOR POLYTOMOUS ATTRIBUTE DATA

Item	P(Obs >= Rep)
1	0.319
2	0.425
3	0.444
4	0.557
5	0.495
6	0.438
7	0.346
8	0.352
9	0.400
10	0.421
11	0.416
12	0.444
13	0.431
14	0.439
15	0.561
16	0.348
17	0.462
18	0.386
19	0.455
20	0.411
21	0.210
22	0.384
23	0.376
24	0.385
25	0.446
26	0.549
27	0.176
28	0.343
29	0.393
30	0.458

Note. The item fit criterion is that the *P* value is around .5 and is not larger than .95 or smaller than .05.

TABLE XXX DESCRIPTIVE STATISTICS FOR PERSON FIT INDICES

Statistics	Value
Mean	0.484
Median	0.481
Standard Deviation	0.255
Minimum	0.006
Maximum	0.988

Note. The person fit criterion is that the *P* value is around .5 and is not larger than .95 or smaller than .05.

TABLE XXXI GROUP-LEVEL CLASSIFICATION RESULTS

			Attributes		
Mastery	1	2	3	4	5
Non-mastery	28.50%	27.30%	20.80%	30.40%	50.10%
Medium mastery	52.00%	50.00%	56.10%	50.60%	49.90%
High mastery	19.50%	22.70%	23.10%	19.00%	0.00%

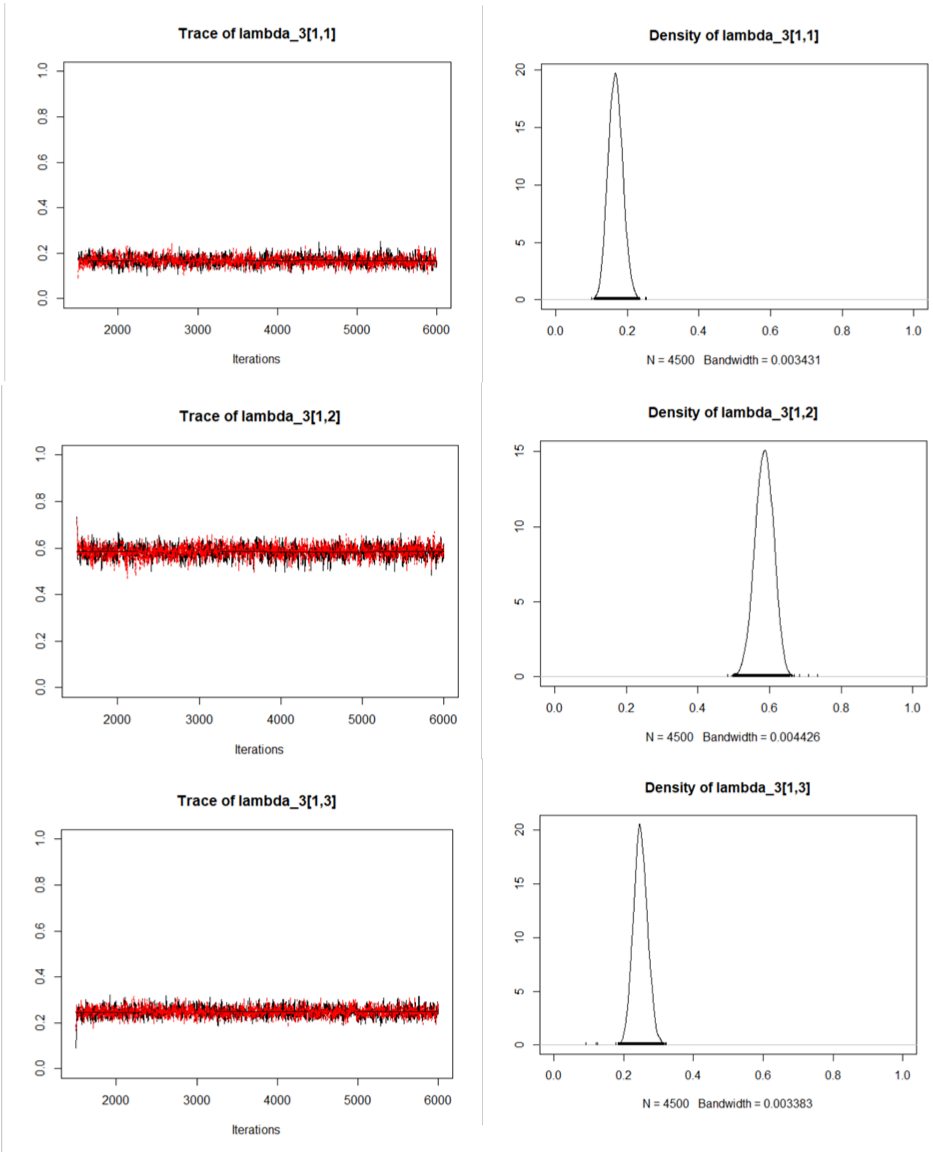


Figure 29. MCMC chain plots and density plots for parameters associated with λ_3.

Note. The value in the bracket [b, a] denotes the *a*th parameter associated with the λ parameter.

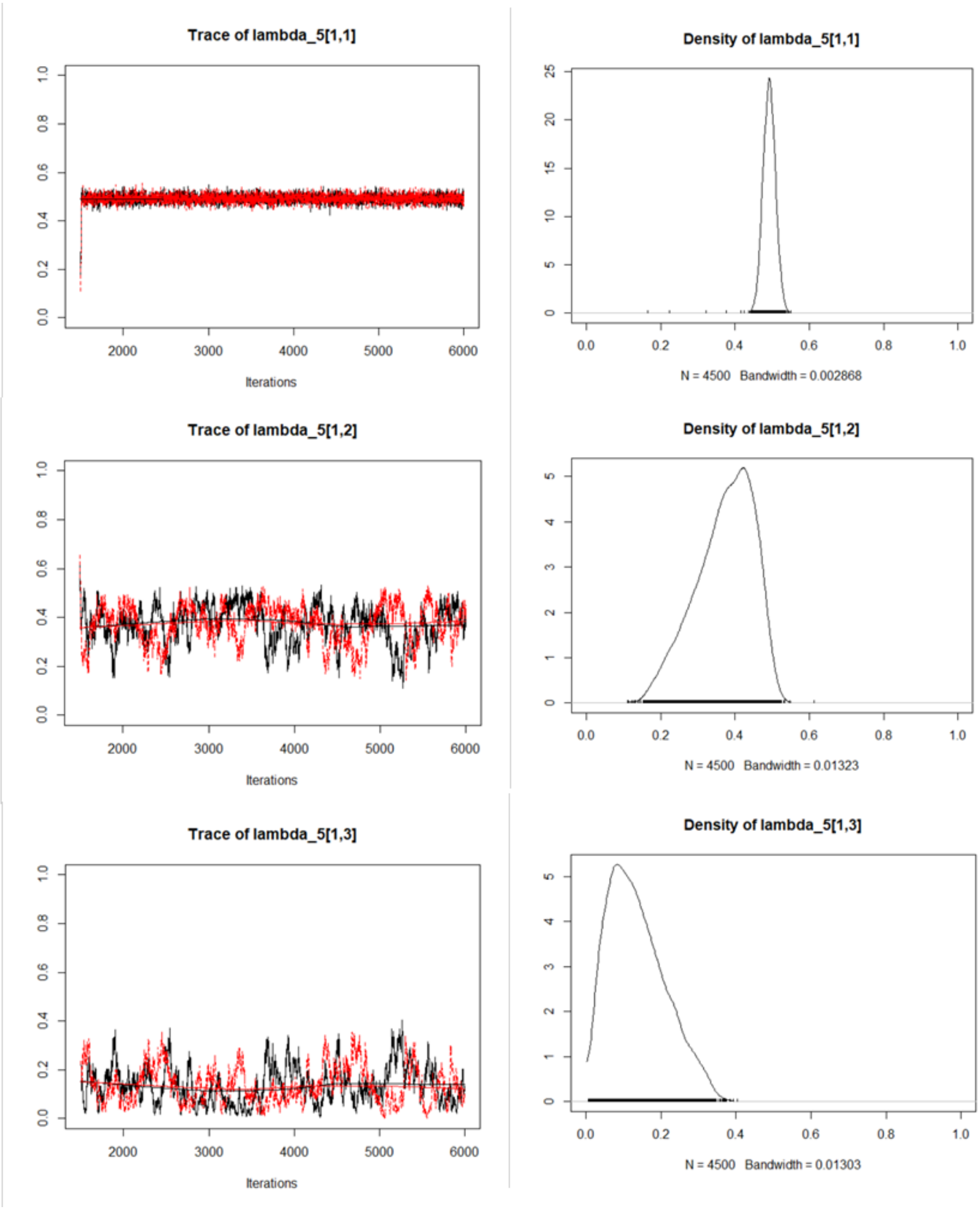


Figure 30. MCMC chain plots and density plots for parameters associated with λ_5.

Note. The value in the bracket [b, a] denotes the *a*th parameter associated with the λ parameter.

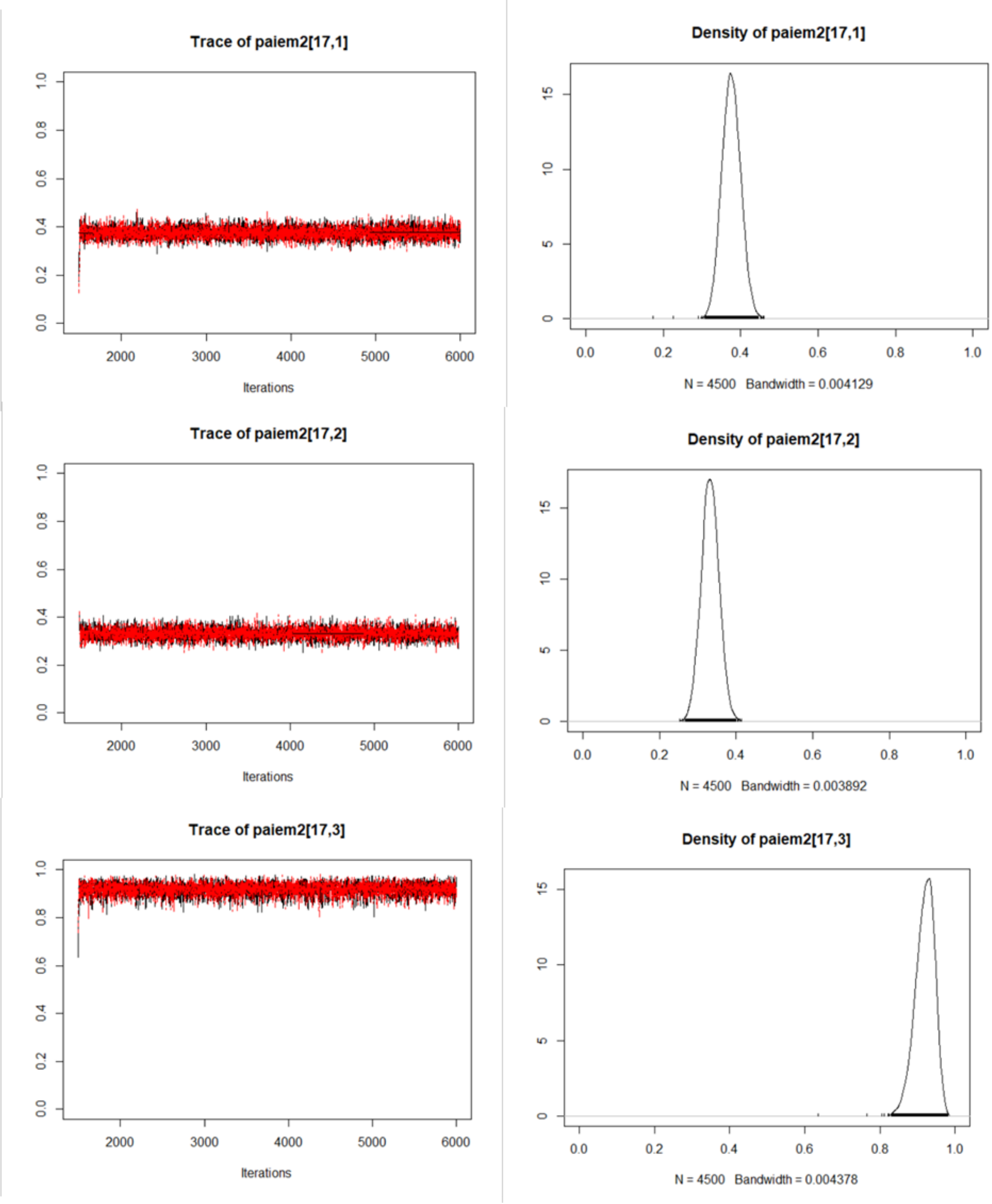


Figure 31. MCMC chain plots and density plots for parameters associated with π_{17}.

Note. The value in the bracket [b, a] denotes the *a*th parameter associated with the π_b parameter.

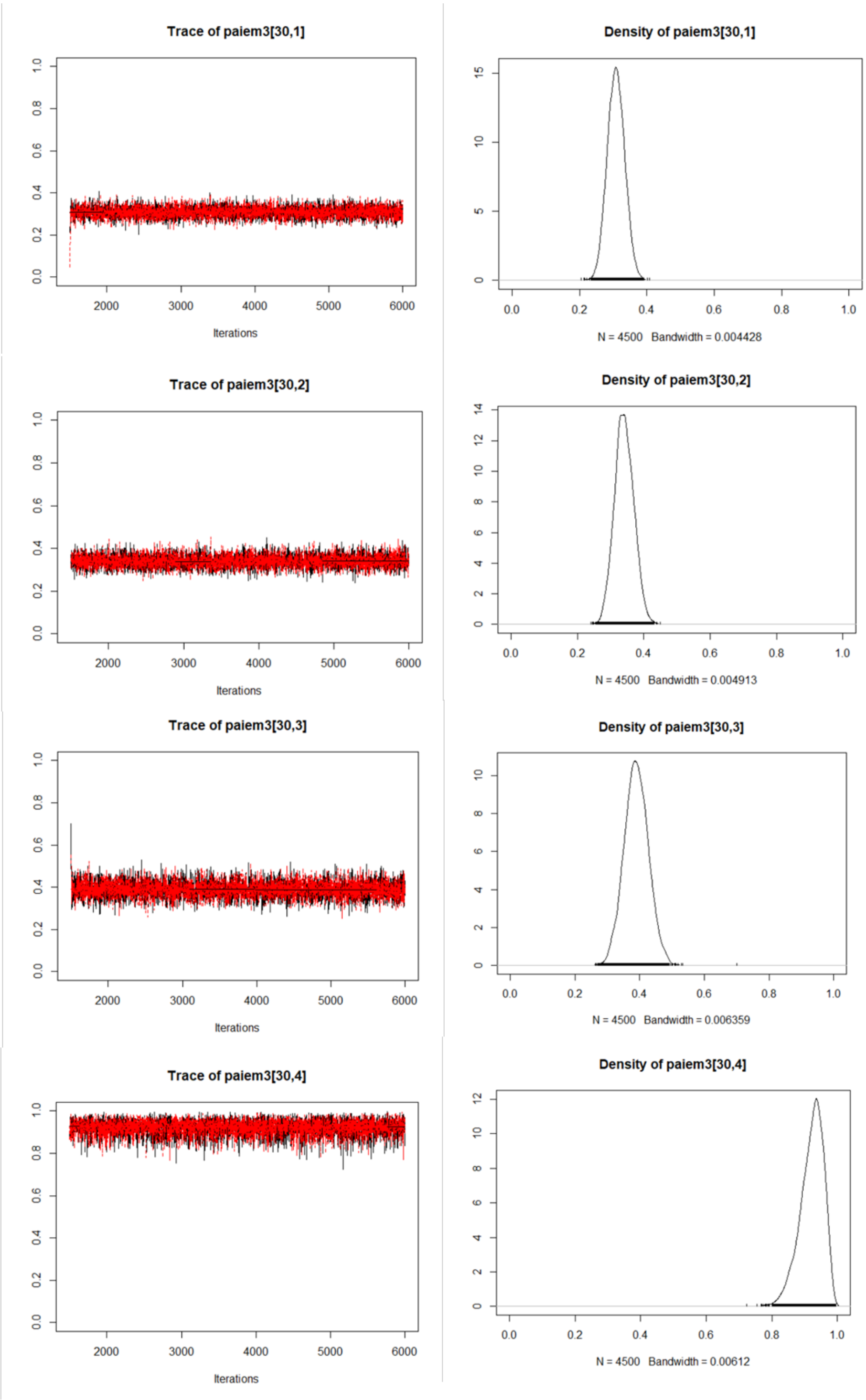
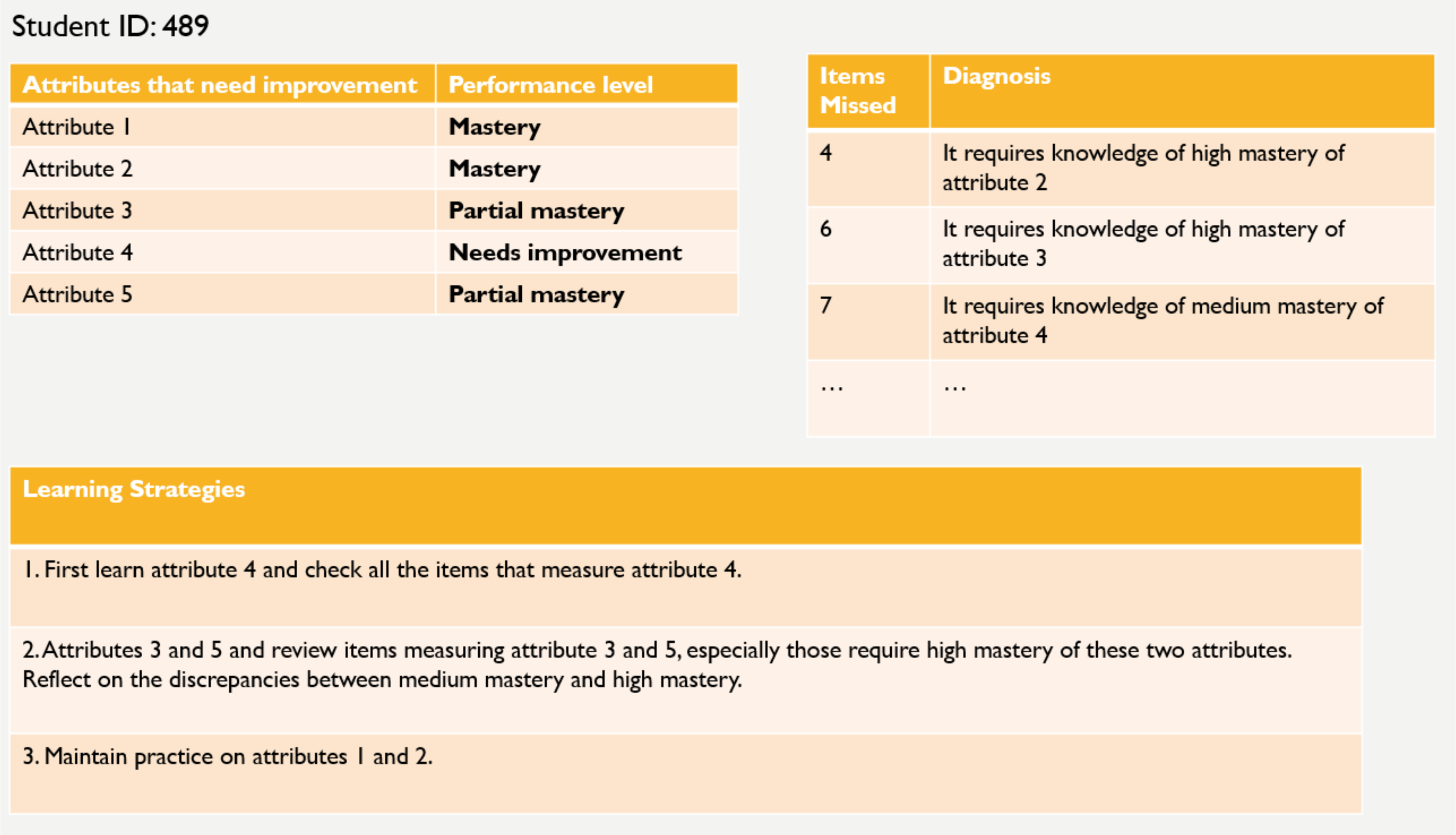


Figure 32. MCMC chain plots and density plots for parameters associated with π_{30}.

Note. The value in the bracket [b, a] denotes the *a*th parameter associated with the π_b parameter.



Student ID: 489

Attributes that need improvement	Performance level
Attribute 1	**Mastery**
Attribute 2	**Mastery**
Attribute 3	**Partial mastery**
Attribute 4	**Needs improvement**
Attribute 5	**Partial mastery**

Items Missed	Diagnosis
4	It requires knowledge of high mastery of attribute 2
6	It requires knowledge of high mastery of attribute 3
7	It requires knowledge of medium mastery of attribute 4
…	…

Learning Strategies
1. First learn attribute 4 and check all the items that measure attribute 4.
2. Attributes 3 and 5 and review items measuring attribute 3 and 5, especially those require high mastery of these two attributes. Reflect on the discrepancies between medium mastery and high mastery.
3. Maintain practice on attributes 1 and 2.

Figure 33. Diagnostic report for Student 489.

5. DISCUSSION

To promote instruction and learning, we need to explore the theoretical and statistical approaches that can well reflect instructionally relevant and cognitively diagnostic assessment results. Compared to other multidimensional classification models, BN alleviates most modeling constraints and supports a complex modeling when there is a need to understand the underlying process of student knowledge. A limited number of studies provided informative feedback regarding students' attribute proficiency and unpacking the structural relationships among attributes, therefore it is necessary to examine whether the BN approach has the capacity to provide diagnostic information on student attribute mastery based on a simulation study and real data illustration. Further, the incorporation of ECD in BN allows the integration between the theoretical foundation of assessment and the modeling technique of psychometrics. This integration would provide practical implications to help teachers customize their instruction to students with different learning needs and to assist students to prioritize their learning goals to make improvement.

This study examines the capacity of BN modeling for diagnosing student attribute mastery and addresses questions related to the analysis and interpretation of assessment data for a formative diagnostic purpose. First, it investigates the effects of sample size, test length, Q matrix complexity, and attribute type on measurement quality. It then compares the performance of BN and GDINA in analyzing assessments under different scenarios and different levels of prior information on attribute structure. Finally, it applies BN into two existing datasets. This chapter presents conclusions based on the results of research questions. It also discusses the limitations of this study and implications for further research.

5.1 Conclusions

This section revisits and draws conclusions on the results presented in Chapter 4.

First, the simulation study in general shows that BN modeling can recover parameter values with small biases and *RMSD*s across different conditions of sample size and test length . Although small, the biases between conditions are slightly different from each other. The estimation and classification accuracy slightly improve as test lengths and sample sizes increase. It is understandable that larger sample sizes and longer test length would provide more information for model estimation. However, this difference across conditions does not preclude the application of BN modeling in analyzing small sample size and short test length. The small magnitude of the difference demonstrates the acceptable stability and accuracy in BN parameter estimation and attribute classification among different assessment contexts. As a result, the recovery of parameters is adequate and satisfactory for both the item parameters under the evidence model and the person parameters under the proficiency model, when BN is used on various conditions of sample size and test length.

Second, regarding the effects of Q matrix complexity and attribute types, it is found that the estimation and classification accuracy tends to be higher for simple structure Q matrix than the complex structure. Further, the analysis of BN modeling on the dichotomous attribute assessment data would yield slightly higher accuracy than the polytomous attribute assessment data. This improvement in simpler models might be attributable to the less complicated relationships among attributes and items and the more straightforward specification of attributes. It is understandable that the reduction in complex information would help models to yield more accurate estimation. However, it does not mean that BN only fits assessments with a simple design. Instead, the rather small difference in estimation accuracy between simple and complex conditions manifest the flexibility of BN modeling in analyzing assessments of different types.

Third, this study examines the classification accuracy of student mastery diagnosis after BN is provided with different amount of information on the attribute structure. Intuitively, more information would bring more accurate classifications across all conditions of Q matrix complexity and attribute types. However, the results of BN modeling with partial information or wrong information still maintain acceptable mastery classification accuracies. When no information is provided, the classification accuracies tend to exhibit larger discrepancy from the true classifications. The graphical display of the relationships among attributes and items allow us to dismantle how the amount of prior information impacts the estimation on mastery probability of each attribute and correctness probability of each item. As it turns out, the lack of some information or wrong information on attribute structure would only influence the classification results of few students. It further supports that BN is robust to yield attribute classifications even with some but not all information on attribute structure.

Fourth, the comparison with CDM evaluated the performance of the two approaches in analyzing the same datasets across different conditions of Q matrix complexity and attribute types. BN modeling tends to perform better in model fitting and produce more accurate attribute mastery classifications across all conditions. Moreover, BN modeling is more capable of incorporating attribute structure into estimation by considering all possible mastery patterns, providing flexibility in modeling. One limitation of CDM is its reduction in potential latent classes after accounting for attribute structure. Another disadvantage of using CDM is its incompatibility with the integration of attribute structure in the polytomous-attribute assessment.

Teachers or researchers may render it questionable to continue using CDM when designing a complex assessment with both a structure in attributes and polytomous attributes.

Finally, the analyses of two existing datasets of dichotomous attributes and polytomous attributes show that BN modeling yield good model fit and parameter fit including item and person parameters. It can also help generate detailed diagnostic reports for each student that would contribute to their remediation learning paths. The graphical display of the relationships between attributes and items would further help teachers and researchers to disentangle the structure in content knowledge and glean an impression of individual-level and group-level performance on each attribute.

5.2 Limitations

Nevertheless, this study has several limitations.

First, this study included a limited variety in attribute structure. For the simulation study design, I used the relationships among attributes proposed by Sinharay, Almond, and Yan (2004) because I intended to simulate an assessment that has an existing theoretical rationale and demonstrates a validated attribute structure and a validated Q matrix. In the TIMSS data analysis, the relationships among a variety of attributes are structural. However, its test length is relatively short, therefore not every attribute is measured by enough number of items. Future research may explore the performance of the BN approach for assessments of more complicated structural relationships among attributes with an adequate test length.

Second, there is a lack of real data of assessments designed for different scenarios (i.e., different levels of Q matrix complexity, different attribute types, etc.). In this case, the interpretation of the results in this study is limited to simulation scenarios instead of practical contexts. Specifically, without a real situation, the interpretation of item and person parameters are out of context and may not furnish meaningful reflections on other components of the ECD structure: for instance, how the current assembly rules of items can reflect a fair amount assessment for each attribute and how the evidence model can provide feedback to assessment design.

Third, the TIMSS data has a sparse Q matrix, in which 23 items were used to measure 13 attributes. It seems ineffective to measure a variety of attributes in a test with limited items as each attribute cannot be fully evaluated by enough items. Therefore, future research may consider narrowing down or collapsing some attributes in a sparse matrix or increasing test length in an effort to furnish enough information on each attribute for the measurement model to make reliable conclusions.

Fourth, this study examines one scenario in the conditions of partial information and wrong information, respectively. Different levels of partial or wrong information may impact the estimation accuracy of BN differently. In practice, teachers of different grades and classes and educational researchers may all have various opinions on the attribute structure. Given the potential disparity that may occur in the attribute structure, future research can explore how different structures in attributes may affect the BN modeling results with more nuance.

Finally, the application of the GDINA model was undertaken on the BN-generated data. However, BN and GDINA are built under the same framework of latent class models so that the data generated from both models are meant to identify the latent groups each person may belong to. In this case, the BN data should be compatible with the application of GDINA, and the deviances of GDINA in this study showed small differences from those of BN, further indicating that there should be small difference in the nature of the two approaches. For an exploratory purpose, future research may use the data generated by each model to evaluate their performance.

5.3 Contributions of This Study and Implications for Future Research

This study makes several practical contributions to the field of assessment and psychometrics through an examination of BN in making formative diagnosis of student performance. This study highlights the potential benefits of BN and its integration in measurement models when classroom assessments are analyzed for diagnostic purposes. In this section, I discuss the contributions of this study and suggested future direction for research.

The natural demonstration of ECD in BN modeling builds a bridge between the underlying theory of the assessed subject domain and the psychometric modeling of assessment data. The development of the proficiency model and the evidence model in ECD relies on theory and expert opinion regarding the structure in attributes and the specification of Q matrix. The incorporation of expert experience helps to emphasize the role of subject matter experts' information in the psychometric analysis of assessment data rather than the mere dependence on the observed data. As such, ECD contributes to BN modeling by echoing the substantive theory of a subject domain in parameter estimation, which further lends support to the interpretability of each parameter under the guidance of the proficiency model and the evidence model. Taken together, the BN approach emphasizes the co-acting influences of conceptual framework of the learning structure in content knowledge guided by ECD and empirical operationalization of measurement models in making formative diagnosis of student performance.

This study demonstrates that BN can accommodate assessments with small sample size and short test length. For classroom assessments, there is a tradeoff between test scale and analysis accuracy. For instance, it is often infeasible to administer formative diagnostic assessments among a large number of students or administer a very long test, while typical psychometric analysis needs more students and longer tests for estimation accuracy. This study examines BN as a possible solution to alleviate this dilemma and reveals the utility of BN in producing results with acceptable accuracy for small-scale formative diagnostic assessments. Note that diagnostic classroom assessments, to some degree, can tolerate less accurate results of parameter estimation and attribute classification compared to high-stakes standardized testing. As such, BN opens the door for the expansion of psychometric analysis to small-scale classroom assessments, which are usually analyzed by raw scores.

BN can also handle both simple and complex assessments with acceptable accuracy. Although complex assessment types (i.e., complex Q matrix structure and polytomous attributes) may lead to less accurate estimation than simple types, they play an important role in reflecting more formative and practical feedback on student knowledge of applying multiple attributes in problem-solving and on finer attribute mastery classifications. The application of BN helps to maintain a balance between estimation accuracy and assessment complexity by rendering reliable results with acceptable error rates.

BN appears to be a useful modeling approach that can reveal student performance levels at both individual and group levels. BN can reveal the individualized gaps between students' actual and expected performance levels and inform students and teachers with suggestions on how to improve their performance. Specifically, through the BN approach, the estimated distribution of mastery probabilities for attributes and the correctness probabilities for items would reveal the concepts and knowledge components that are difficult for students to understand and that students have mastered well. Moreover, understanding student performance levels may yield insights into ways of identifying the causes of knowledge gaps as well as the strengths and weaknesses of a student’s understanding toward a given concept. This information would provide students with remediation learning paths and teachers with instructional plans customized for students’ needs.

As discussed above, BN modeling relies on the ECD framework. The BN model results may also in turn provide feedback to future test design and the test assembly model in the ECD framework. For formative assessment purpose, teachers may use the latest BN diagnostic information to appraise whether the results are consistent with their assessment purposes and expectations. For example, the BN diagnostic results may indicate that students have not established the structure among some attributes. In the next formative assessment, teachers can choose to reinforce the structure and examine whether students improve their understanding after taking the last assessment. That is, teachers and researchers may explore better ways to teach and a more robust way to assess attributes at both group and individual levels.

Future research on analyzing formative diagnostic assessment data should endeavor to fully develop the potentials of graphical modeling and flexibility of BN for drawing inferences about student understanding of domain knowledge. The posterior distribution over attributes and items in the BN graph can demonstrate a customized and straightforward diagnosis on how students may perform under different mastery profiles and how attributes are influenced by each other. Such information can help teachers to increase knowledge on effective instructional materials for students’ remediation on the attributes they have not mastered and make decisions regarding the development of new diagnostic assessments.

Another topic on the diagnostic assessment that future research might delve into is the identification of Q matrix. It is unclear whether BN can identify or correct the mis-specified Q matrices. That is, under the situation where the Q matrix partially or wrongly identified the measured attributes for each item, it is necessary to explore whether the BN approach can evaluate the quality of Q matrix and further identify the mis-specified items through the observed item responses and the relationships among attributes.

On top of the potential issues with Q matrix identification, the assumption on the attribute structure might be inaccurate as well. As demonstrated in the results, the group-level diagnostic report on each attribute may allow teachers to identify any inconsistencies between the group-level mastery of attributes and the structure in attributes. These inconsistencies would further guide the revision of the hypothesized attribute structure and the decision on the numbers of items measuring each attribute in a test.

The flexibility of BN also reveals in its capacity of accounting for covariates. In many cases, assessment contexts may differ across classrooms and student of different backgrounds may perform diversely. Therefore, in the BN modeling, researchers can include covariates that possibly impact the cognitive diagnostic results and differ across students in order to better individualize assessments. For example, the covariate of students' performance on previous classroom tests can be added to better estimate student ability level and understand students' learning trajectories.

To increase our confidence in the diagnostic results of the BN approach, we can apply statistical cross-validation and field-test validity study to determine the performance of the BN approach on additional data or contexts. Specifically, statistical cross-validation methods include holdout cross-validation, k-fold cross-validation, leave-one-subject-out cross-validation, leave-one-trial-out cross-validation, etc. (Lever et al., 2016). The choice of the cross-validation method depends on modeling factors including the sample size and the research design (Koul et al., 2018). In addition, a validity study can be conducted to decide whether the cognitive diagnostic results generated by the BN approach are consistent with those obtained from a similar test (Liu et al., 2013). Furthermore, for an application of BN in a real test, we could interview teachers about whether the BN diagnostic results are consistent with their estimation of students' understanding of attributes.

Although BN modeling performs well in recovering parameter values and attribute classifications, the Bayesian inferences may rely on the assumptions about the relationships among attributes, the specification of Q matrix, and the starting values of parameters (i.e., priors). In other words, especially under the scenario of a small sample size and a short test length, which reveals less information provided from the observed data, the Bayesian estimation tends to rely more on the predetermined prior information. The subjectivity induced by the choice of priors are the features but sometimes disadvantages of the Bayesian approach. Therefore, a refined Q matrix and a relatively accurate estimation for priors are recommended to achieve a satisfactory estimation accuracy of the BN results, especially in small-scale assessments. This information would require substantial theoretical and empirical work. To scale up the use of the BN approach, ideally teachers or test developers may discuss with subject matter experts on what attributes are measured by each item and what percent of students are expected to master an attribute among the target population.

In conclusion, BN serves as a flexible modeling approach that can accommodate various assessment scenarios. It is robust in making satisfactory and stable parameter estimation and student mastery classification for small-scale classroom assessments. This study highlights a practical consideration for the application of BN in cognitively diagnosing student mastery of knowledge and graphically displaying the relationships between attribute structure and item responses. More studies may be conducted to further explore this promising method in the future.

REFERENCES

Almond, R. G., DiBello, L. V., Moulder, B., & Zapata-Rivera, J.-D. (2007). Modeling diagnostic assessments with Bayesian networks. *Journal of Educational Measurement, 44*(4), 341-359.

Almond, R. G., Mislevy, R. J., Steinberg, L. S., Yan, D., & Williamson, D. M. (2015). *Bayesian networks in educational assessment*. New York, NY: Springer.

Almond, R. G., Shute, V. J., Underwood, J. S., & Zapata-Rivera, J.-D. (2009). Bayesian networks: A teacher's view. *International Journal of Approximate Reasoning, 50*(3), 450-460.

Bennett, R. E. (2018). Educational assessment: What to watch in a rapidly changing world. *Educational Measurement: Issues and Practice, 37*(4), 7-15.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7-74.

Black, P., & Wiliam, D. (2005). *Inside the black box: Raising standards through classroom assessment*. London, UK: Granada Learning.

Bolt, D. (2007). The present and future of IRT-based cognitive diagnostic models (ICDMs) and related methods. *Journal of Educational Measurement*, *44*(4), 377-383.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434-455.

Chen, J., & de la Torre, J. (2013). A general cognitive diagnosis model for expert-defined polytomous attributes. *Applied Psychological Measurement*, *37*(6), 419-437.

Chen, Y., Li, X., Liu, J., & Ying, Z. (2018). Recommendation System for Adaptive Learning. *Applied Psychological Measurement, 42*(1), 24-41.

Chiu, C.-Y., Sun, Y., & Bian, Y. (2018). Cognitive diagnosis for small educational programs: The general nonparametric classification method. *Psychometrika, 83*(2), 355-375.

Choi, Y. (2012). *Dynamic bayesian inference networks and hidden markov models for modeling learning progressions over multiple time points.* (Doctoral Dissertation), Retrieved from https://drum.lib.umd.edu/handle/1903/12739

Culbertson, M. (2016). Bayesian networks in educational assessment: The state of the field. *Applied Psychological Measurement, 40*(1), 3-21.

Culbertson, M. J. (2014). *Graphical models for student knowledge: Networks, parameters, and item selection.* (Doctoral dissertation), Retrieved from http://hdl.handle.net/2142/49372

Culbertson, M. J., & Li, F. (2012). *Analyzing knowledge structure: An application of graphical models to a medical licensure exam*. Paper presented at the annual meeting of National Council on Measurement in Education, Vancouver, Canada.

De La Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179-199.

De La Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología Educativa, 20*(2), 89-97.

Desmarais, M. C., & Pu, X. (2005). A Bayesian student model without hidden nodes and its comparison with item response theory. *International Journal of Artificial Intelligence in Education, 15*(4), 291-323.

DiBello, L. V., Roussos, L. A., & Stout, W. (2006). 31A Review of Cognitively Diagnostic Assessment and a Summary of Psychometric Models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics* (Vol. 26, pp. 979-1030): Elsevier.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Feinberg, R. A., & Wainer, H. (2014). When can we improve subscores by making them shorter? : The case against subscores with overlapping items. *Educational Measurement: Issues and Practice, 33*(3), 47-54.

Haberman, S., Sinharay, S., & Puhan, G. (2009). Reporting subscores for institutions. *British Journal of Mathematical and Statistical Psychology, 62*(1), 79-95.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*(4), 301-321.

Hashimoto, T., & Ueno, M. (2011). Latent conditional independence test using Bayesian network item response theory. *IEICE TRANSACTIONS on Information and Systems, 94*(4), 743-753.

Huff, K., Steinberg, L., & Matts, T. (2010). The promises and challenges of implementing evidence-centered design in large-scale assessment. *Applied Measurement in Education, 23*(4), 310-324.

Im, S., & Yin, Y. (2009). Diagnosing skills of statistical hypothesis testing using the rule space method. *Studies in Educational Evaluation, 35*(4), 193-199.

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to LanguEdge assessment. *Language Testing, 26*(1), 031-073.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258-272.

Karabatsos, G. (2017). A menu-driven software package of Bayesian nonparametric (and parametric) mixed models for regression analysis and density estimation. *Behavior Research Methods*, *49*(1), 335-362.

Karabatsos, G., & Sheu, C.-F. (2004). Order-constrained Bayes inference for dichotomous models of unidimensional nonparametric IRT. *Applied Psychological Measurement, 28*(2), 110-125.

Khajah, M. M., Huang, Y., González-Brenes, J. P., Mozer, M. C., & Brusilovsky, P. (2014). *Integrating knowledge tracing and item response theory: A tale of two frameworks.* Paper presented at the CEUR Workshop Proceedings.

Koller, D., Friedman, N., & Bach, F. (2009). *Probabilistic graphical models: Principles and techniques*. Cambridge, MA: MIT press.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174.

Lee, H.-S., Gweon, G.-H., Dorsey, C., Tinker, R., Finzer, W., Damelin, D., Kimball, N., Pallant, A., Lord, T. (2015). *How does Bayesian knowledge tracing model emergence of knowledge about a mechanical system?* In *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge* (pp. 171–175).

Lee, Y.-S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the US national sample using the TIMSS 2007. *International Journal of Testing, 11*(2), 144-177.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement, 41*(3), 205-237.

Lever, J., Krzywinski, M., & Altman, N. (2016). Model selection and overfitting. Vol. 13. *Nature Methods*, 703-4.

Levy, R., & Mislevy, R. J. (2004). *Specifying and Refining a Measurement Model for a Simulation-Based Assessment.* . Retrieved from https://eric.ed.gov/?id=ED483385

Liu, C. L. (2009). Selecting Baysian-network models based on simulated expectation. *Behaviormetrika, 36*(1), 1-25.

Liu, H. Y., You, X. F., Wang, W. Y., Ding, S. L., & Chang, H. H. (2013). The development of computerized adaptive testing with cognitive diagnosis for an English achievement test in China. *Journal of Classification*, *30*(2), 152-172.

Meyn, S. P., & Tweedie, R. L. (2012). *Markov chains and stochastic stability*. London, UK: Springer Science & Business Media.

Mislevy, R. J., Almond, R. G., Yan, D., & Steinberg, L. S. (1999). *Bayes nets in educational assessment: Where the numbers come from.* In *Proceedings of the fifteenth conference on uncertainty in artificial intelligence* (pp. 437–446).

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice, 25*(4), 6-20.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3-62.

Norsys. (1992-2014). Netica manual. Vancouver, BC.

Nouh, Y., Karthikeyani, P., & Nadarajan, R. (2006). Updating Student Model using Bayesian Network and Item Response Theory. In *2006 Fourth International Conference on Intelligent Sensing and Information Processing* (pp. 161-164). IEEE.

Park, Y. S., & Lee, Y.-S. (2014). An extension of the DINA model using covariates: Examining factors affecting response probability and latent classification. *Applied Psychological Measurement, 38*(5), 376-390.

Pek, P.-K., & Poh, K.-L. (2004). A Bayesian tutoring system for Newtonian mechanics: Can it adapt to different learners? *Journal of Educational Computing Research, 31*(3), 281-307.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Plummer, M. (2015). JAGS Version 4.0. 0 user manual. Retrieved from http://sourceforge. net/projects/mcmc-jags

Puhan, G., Sinharay, S., Haberman, S., & Larkin, K. (2010). The utility of augmented subscores in a licensure exam: An evaluation of methods using empirical data. *Applied Measurement in Education*, *23*(3), 266-285.

Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. S. (2010). *Handbook of educational data mining*. Boca Raton, FL: CRC press.

Rupp, A. (2002). Feature Selection for Choosing and Assembling Measurement Models: A Building-Block-Based Organization. *International Journal of Testing, 2*(3-4), 311-360.

Rupp, A., Henson, R., & Templin, J. (2010). *Diagnostic Measurement : Theory, Methods, and Applications*. New York, NY: Guilford Publications.

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement, 6*(4), 219-262.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. *Computer Games and Instruction*, *55*(2), 503-524.

Sinharay, S., Almond, R., & Yan, D. (2004). Assessing fit of models with discrete proficiency variables in educational assessment. *ETS Research Report Series, 2004*(1), i-49.

Steedle, J. T. (2008). *Latent class analysis of diagnostic science assessment data using Bayesian networks.* (Doctoral Dissertation), PsycINFO database.

Su, Y.-L., Choi, K., Lee, W., Choi, T., & McAninch, M. (2013). Hierarchical cognitive diagnostic analysis for TIMSS 2003 mathematics. *Centre for Advanced Studies in Measurement and Assessment, 35*, 1-71.

Su, Y.-S., Yajima, M., & Su, Y.-S. (2015). Package ‘R2jags’. Retrieved from http://CRAN. R-project. org/package= R2jags

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*(4), 345-354.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. *Cognitively Diagnostic Assessment*, 327-359.

Tatsuoka, K. K., & Tatsuoka, M. M. (1992). A psychometrically sound cognitive diagnostic model: Effect of remediation as empirical validity. *ETS Research Report Series, 1992*(1), i-24.

Templin, J., & Bradshaw, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika, 79*(2), 317-339.

Templin, J. L. (2004). *Generalized linear mixed proficiency models.* (Doctoral dissertation), Retrieved from https://www.ideals.illinois.edu/handle/2142/82077

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287-305.

Tu, D., Wang, S., Cai, Y., Douglas, J., & Chang, H.-H. (2019). Cognitive diagnostic models with attribute hierarchies: Model estimation with a restricted Q-matrix design. *Applied Psychological Measurement, 43*(4), 255-271.

Ueno, M. (2002). An extension of the IRT to a network model. *Behaviormetrika, 29*(1), 59-79.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*(2), 287-307.

West, P., Rutstein, D. W., Mislevy, R. J., Liu, J., Levy, R., Dicerbo, K. E., Crawford, A., Choi, Y., Chapple, K. & Behrens, J. T. (2012). A Bayesian Network Approach to Modeling Learning Progressions. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning Progressions in Science: Current Challenges and Future Directions* (pp. 257-292). Rotterdam: SensePublishers.

Wilson, K. H., Karklin, Y., Han, B., & Ekanadham, C. (2016). Back to the basics: Bayesian extensions of IRT outperform neural networks for proficiency estimation. *Proceedings of the 9th International Conference on Educational Data Mining.*

Wu, H. (2014). *A comparison of General Diagnostic Models (GDM) and Bayesian networks using a middle school mathematics test.* (Doctoral dissertation), PsycINFO database.

Xenos, M. (2004). Prediction and assessment of student behaviour in open and distance education in computers using Bayesian networks. *Computers & Education, 43*(4), 345-359.

Yan, D., Mislevy, R. J., & Almond, R. G. (2003). Design and analysis in a cognitive assessment. *ETS Research Report Series*, *2003*(2), i-47.

Yin, Y., Tomita, M. K., & Shavelson, R. J. (2014). Using formal embedded formative assessments aligned with a short-term learning progression to promote conceptual change and achievement in science. *International Journal of Science Education, 36*(4), 531-552.

Zhang, S., & Chang, H.-H. (2016). From smart testing to smart learning: How testing technology can assist the new generation of education. *International Journal of Smart Technology and Learning, 1*(1), 67-92.

APPENDIX

Notice of Determination
Activity Does Not Represent Human Subjects Research

January 7, 2020

Xiaodan Tang, MA
Educational Psychology
Phone: (734) 545-9783

RE: **Protocol # 2020-0024**
“Graphically Modeling Student Knowledge: A Bayesian Network Approach”

Sponsor: None

Dear Mx. Tang:

The UIC Office for the Protection of Research Subjects received your application, and has determined that this activity **DOES NOT meet the definition of human subject research** as defined by 45 CFR 46.102(e).

Specifically, an analysis of publicly available, de-identified data from the U.S. Trends in International Mathematics and Science Study, 2003 Grade 8 Mathematics test. No restricted-use data from this repository will be used.

You may conduct your activity without further submission to the IRB.

Please note:

- If this activity is used in conjunction with any other research involving human subjects, prospective IRB approval or a Claim of Exemption is required.
- If this activity is altered in such a manner that may result in the activity representing human subject research, a NEW Determination application must be submitted.

Sincerely,
Sandra Costello
Assistant Director, IRB # 7
Office for the Protection of Research Subjects

cc: Stacey S. Horn, Educational Psychology, M/C 147
Yue Yin (faculty advisor), Educational Psychology, M/C 147

VITA
Xiaodan Tang

Phone: 734-545-9783   Email: xtang1322@gmail.com

EDUCATION

University of Illinois at Chicago

Ph.D. in Measurement, Evaluation, Statistics, and Evaluation (MESA) *August 2014 - August 2020*

GPA: 4.00/4

Dissertation: “Graphically Modeling Student Knowledge: A Bayesian Network Approach”

Committee: Dr. Yue Yin (Chair), Dr. George Karabatsos, Dr. Yoon Soo Park, Dr. James Pellegrino, Dr. Hua-Hua Chang

M.S. in Statistics *August 2017 - December 2019*

GPA: 3.88/4

University of Michigan - Ann Arbor

M.A. in Educational Studies *August 2012 - December 2013*

GPA: 3.91/4

Nanjing Normal University

B.A. in Teaching Chinese as a Second Language *September 2008 - June 2012*

Major GPA: 3.90/4.

PUBLICATIONS

Lin, Q., Yin, Y., **Tang, X.,** Hadad, R. & Zhai, X. (2020). A Systematic Review of Research on Assessments Used in Maker Activities. *Computers & Education.*

Tang, X., Karabatsos, G. & Chen, H. (2020). Detecting local dependence: A threshold-autoregressive item response theory (TAR-IRT) approach for polytomous items. *Applied Measurement in Education.*

Tang, X., & Schultz, M. (2020). The effect of repeat simulation-based item exposure. *Practical Assessment, Research and Evaluation,* Vol. 25, Article 3.

Tang, X., Yin, Y., Lin, Q., Hadad, R. & Zhai, X. (2020). Assessing computational thinking: A systematic review of the literature. *Computers & Education,* 148.

Yin, Y., Hadad, R., **Tang, X.,** & Lin, Q. (2020). Improving and assessing computational thinking in maker activities: the integration with physics and engineering learning. *Journal of Science Education and Technology*, 1-26.

TEACHING EXPERIENCE

Department of Educational Psychology, University of Illinois at Chicago

EPSY 546 Educational Measurement	Instructor (4.4/5.0)	*Spring 2019*
EPSY 547 Multiple Regression	Instructor (4.2/5.0)	*Fall 2018*
EPSY 563 Advanced Analysis of Variance	Instructor (4.3/5.0)	*Spring 2017*
EPSY 405 Educational Evaluation and Assessment	Instructor (4.1/5.0)	*Fall 2016*
ED/EPSY 503 Quantitative Inquiry in Education	Teaching Assistant	*Spring 2016* *& Fall 2015*

Department of Asian Language and Cultures, University of Michigan

ASIANLAN 202 Chinese Language (1)	Teaching Assistant	*Fall 2013*
ASIANLAN 201 Chinese Language (2)	Teaching Assistant	*Spring 2013*

PROFESSIONAL EXPERIENCE

Kaplan Test Prep (KTP) *Summer 2019*

Psychometric Intern

- Conducted item calibration and examined subdomain classification using R based on cognitive diagnostic modeling (CDM) for a licensure medical-related summative computer adaptive practice test (CAT), wrote a proposal being accepted by NCME
- Engaged in training on measurement and operational work topics such as assessment platform, data cleaning, and item analysis

American Institute of Certified Public Accountants (AICPA) *Summer 2019*

Psychometric Intern

- Examined the impact of repeat simulation-based item exposure on test taker performance in a multistage (MST) licensure test
- Wrote a proposal being accepted by NCME, and engaged in training on exam structure and scoring, automated test assembly (ATA), operational assessment procedures

National Commission on Certification of Physician Assistants (NCCPA) *Summer 2018*

Psychometric Intern

- Wrote R code to simulate a complex continuous medical-related certification exam, developed a feedback mechanism with survey questions, compared scoring methods, wrote a proposal and presented at NCME 2019
- Examined practical issues in innovative item scoring and longitudinal assessment

Measurement, Evaluation, Statistics, Assessment Lab, University of Illinois at Chicago *August 2017- May 2018*

Statistical Consultant

- Provided statistical and psychometric consulting services for faculty and students including topics of data cleaning, survey development and analysis, CTT and IRT analyses, CAT and CAT simulation, item analysis, ANOVA, regression, multilevel modeling, non-parametric modeling, SEM, missing data analysis

RESEARCH EXPERIENCE

Assessing Computational Thinking in Maker Activities (Funded by National Science Foundation), University of Illinois at Chicago. *October 2015 - June 2018*

Graduate Research Assistant

- Conducted literature review on computational thinking assessment
- With the team, co-developed maker activities to improve computational thinking and physics learning
- With the team, co-developed in formative and summative computational thinking assessments
- With the team, collected the implementation data of maker activities
- Collected and analyzed think-alouds data and pre and post test data in R
- Wrote conference proposals and manuscripts

A Longitudinal Study of Placement Stability Over Time for Youth (Funded by Department of Child and Family Services), UIC. *October 2014 - May 2015*

Graduate Research Assistant

- Participated in survey design, wrote technical reports
- Conducted survey data analysis to examine the relationship between foster children's misbehaviors and foster parenting and to predict the placement outcome based on ANOVA, ANCOVA, factor analysis, linear and logistic regression using SPSS and SAS

Studying Inquiry Based Learning in Higher Education, University of Michigan. *December 2013 - August 2014*

Graduate Research Assistant

- Conducted literature review.

The Iterative Development of Modules to Support Teachers' Engagement in Exploring Language and Meaning in Text with English Language Learners, University of Michigan. *September 2013 - December 2013*

Graduate Research Assistant

- Conducted data analysis based on regression, ANOVA, t-test, correlation using STATA, wrote technical reports

CONFERENCE PRESENTATIONS

Tang, X., & Li, C. (2020). *Subdomain Classification for Remediation with a Summative CAT.* Paper accepted at the annual meeting of the National Council on Measurement in Education (NCME), San Francisco.

Tang, X., & Schultz, M. (2020). *The Effect of Repeat Item Exposure in a High-stakes Standardized Assessment.* Paper accepted at the annual meeting of the National Council on Measurement in Education (NCME), San Francisco.

Tang, X., & Chen, H. (2019). *Detection of item parameter drift and item exposure using change point analysis in a CAT context.* Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Toronto.

Tang, X., & Dai, T. (2019). *How do children's social skills influence longitudinal academic achievement? An autoregressive latent trajectory analysis.* Paper presented at the annual meeting of the American Education Research Association (AERA), Toronto.

Tang, X., & Dai, T. (2019). *The role of social skills, behavioral engagement, and gender in early adolescents' academic achievement.* Paper presented at the annual meeting of the American Education Research Association (AERA), Toronto, Canada.

Tang, X., Dallas, D., Fan, F., & Goodman, J. (2019). *A comparison of four IRT-based scoring methods for a continuous assessment.* Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Toronto.

Lin, Q., Yin, Y., **Tang, X.,** & Hadad, R. (2018). *A systematic review of empirical research on maker activity assessments.* Paper presented the annual meeting of at the American Education Research Association (AERA), New York.

Tang, X. (2018). *A multilevel IRT modeling framework for detecting differential item functioning.* Paper presented at the annual meeting of the American Education Research Association (AERA), New York.

Tang, X., Yin, Y., Lin, Q., & Hadad, R. (2018). *Assessing computational thinking: a systematic review of the literature*. Paper presented at the annual meeting of the American Education Research Association (AERA), New York.

Tang, X., Yin, Y., Lin, Q., & Hadad, R. (2018). *Making computational thinking evident: A think-alouds validation study of a computational thinking test.* Paper presented at the annual meeting of the American Education Research Association (AERA), New York.

Tang, X. (2017). *A DIF analysis based on hierarchical generalized linear IRT models*. Paper presented at the the annual meeting of Midwestern Psychological Association (MPA), Chicago, IL.

Tang, X., Karabatsos, G. & Chen, H. (2017). *Modeling local item dependence with an autoregressive IRT method*. Paper presented at the annual meeting of the American Psychology Association (APA), Washington, DC.

Tang, X., Karabatsos, G., & Chen, H. (2017). *Threshold Autoregressive IRT model in polytomous items*. Paper presented at the the annual meeting of National Council on Measurement in Education (NCME), San Antonio, TX.

Tang, X., Yin, Y., Lin, Q., & Hadad, R. (2017). *Assessing computational thinking: a test with a combination of think-aloud and written prompts*. Paper presented at the annual meeting of the American Psychology Association (APA), Washington, DC.

Yin, Y., Hadad, R., **Tang, X.**, Lin, Q., & Hausman, C. M. (2017). *Improving computational thinking skills and physics engineering learning by using makerspace activities and formative assessments.* Paper presented at the annual meeting of the National Association for Research in Science Teaching (NARST), San Antonio, TX.

Tang, X. (2016). *A multilevel IRT model with an application to DIF detection for a teacher evaluation instrument*. Paper presented at the annual meeting of the Mid-Western Educational Research Association (MWERA), Evanston, IL.

QUANTITATIVE SKILLS

Statistical Software	R, SAS, Mplus, HLM, SPSS, Stata, Netica
Measurement Software	flexMIRT, Winsteps, BILOG, Conquest, Parscale, GENOVA, Facets
Other Software	Camtasia, Qualtrics, PollEverywhere, Atlas.ti, Endnote

CERTIFICATION

SAS Certified Advanced Programmer for SAS9	*2018*
TF-CBT (Trauma-Focused Cognitive-Behavioral Therapy) Certificate	*2013*

AWARDS

Student Presenter's Travel Award, University of Illinois at Chicago	*2018*
Graduate Student Travel Award, University of Illinois at Chicago	*2018*
APA Travel Award, American Psychological Association	*2017*
John E. Warriner Scholarship, University of Michigan	*2012*
Chancellor Scholarship, Nanjing Normal University	*2012*
Outstanding Undergraduate Award, Nanjing Normal University	*2012*
Chu Ching-Wen Scholarship, Nanjing Normal University	*2011*

PROFESSIONAL ASSOCIATION MEMBERSHIP

National Council on Measurement in Education	*2015-present*
American Educational Research Association	*2016-present*