Post Facto Analysis of GATE Data: Item Response Theory (IRT) and IRTree Models

A Thesis Submitted

in Partial Fulfillment of the Requirements

for the Degree of

Master of Technology

by

Arijit Ganguly

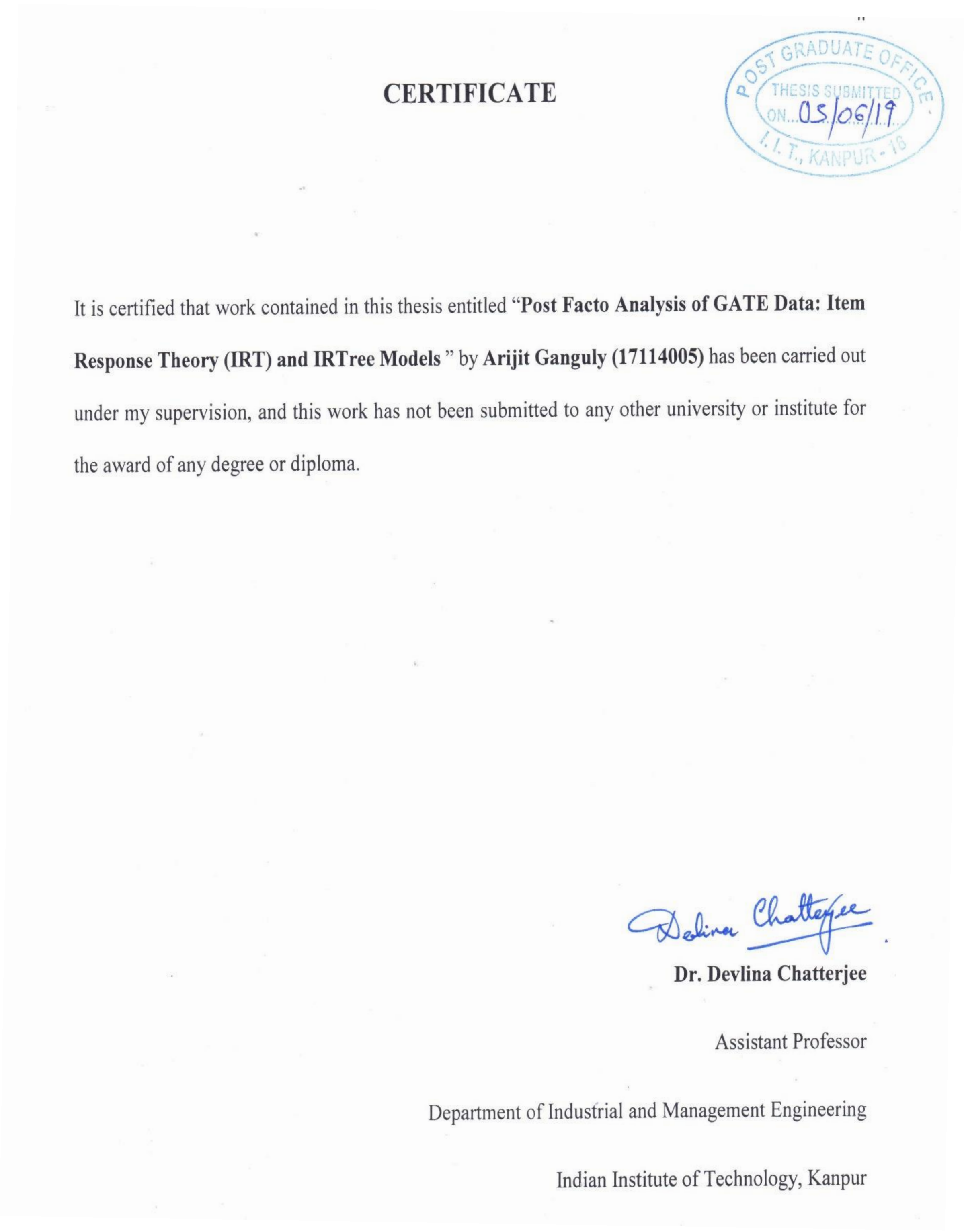
(17114005)

to the

DEPARTMENT OF INDUSTRIAL AND MANAGEMENT ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY KANPUR

June 2019



CERTIFICATE

It is certified that work contained in this thesis entitled "**Post Facto Analysis of GATE Data: Item Response Theory (IRT) and IRTree Models** " by **Arijit Ganguly (17114005)** has been carried out under my supervision, and this work has not been submitted to any other university or institute for the award of any degree or diploma.

Devlina Chatterjee

Dr. Devlina Chatterjee

Assistant Professor

Department of Industrial and Management Engineering

Indian Institute of Technology, Kanpur

Statement of Thesis Preparation

1. Thesis title: **Post Facto Analysis of GATE Data: Item Response Theory (IRT) and IRTree Models.**

2. Degree for which the thesis is submitted: **M.Tech**

3 Thesis Guide was referred to for preparing the thesis.

4. Specifications regarding thesis format have been closely followed.

5. The contents of the thesis have been organized based on the guidelines.

6. The thesis has been prepared without resorting to plagiarism.

7. All sources used have been cited appropriately.

8 The thesis has not been submitted elsewhere for a degree.

Arijit Ganguly.

(Signature of the student)

Name: Arijit Ganguly

Roll No.: 17114005

Department/IDP: IME

Abstract

The Graduate Aptitude Test in Engineering (GATE) is administered annually to assess the scientific and engineering aptitude of bachelor’s degree holders in India. GATE questions are either multiple choice or numerical answer type, with three possible outcomes: omitted, attempted-incorrect, attempted-correct, with a pre-announced scoring formula. There exist alternative methods for evaluation of candidate ability other than traditional formula scoring. Item Response Theory (IRT) is one such method based on statistical modeling of test data, where the probability of getting an answer correct is modeled as a function of both item parameters and the candidate’s latent ability traits. This method has certain advantages over traditional formula scoring methods; for instance it does not constrain the ability estimate of a candidate to be based on pre-determined difficulty parameter of a question as set by the examiner. IRT is widely used in international assessments such as the SAT (Scholastic Assessment Test) and the GRE (Graduate Record Examinations).

In this thesis, we have performed a *post-facto* analysis on the item response dataset of the GATE examination conducted in 2015. We analyze data for 15 out of 23 subjects and our sample consists of 66,084 candidates. Our aim in this work was to identify the models that have the best fit for the GATE 2015 data. Initially, we implemented two variations of traditional IRT models with a single latent trait. Then we implemented four variations of an advanced Tree-based Item response (IRTree) model, which uses a tree structure to replicate the process of attempting a question. Finally, we compare three possible tree structures that represent alternative decision processes of the test taker, even though one of the tree structures is intuitively most pleasing.

Based on three different measures of fit, we found that the linear-tree multidimensional IRTree model having two latent traits (propensity to attempt a question and the ability to answer correctly if attempted) per candidate and two item difficulties per question (intensity of omission induction and easiness) provided the best fit to the dataset for all the 15 subjects. We used the freely available statistical software R and inbuilt packages for the analysis.

We recommend the model with the best fit that can be used to generate ability estimates of candidates, which can be used either as an alternative or as a secondary decision criterion to the current GATE formula score. The models generated by the IRTree methodology in this study provide a statistically robust alternative to traditional formula scoring in estimating candidate ability.

Dedicated to,

My Parents,

Late Mrs. Samita Ganguly

Mr. Abhijit Ganguly

&

My Sister ***Arpita***

Acknowledgment

I express my sincere gratitude to my thesis supervisor, Dr. Devlina Chatterjee, for the invaluable and inspiring guidance, suggestions, and constant encouragement. She provided complete support since the beginning of the work.

I also express my sincere gratitude to Dr. Anindya Chatterjee, Professor at Mechanical Engineering Department, IIT Kanpur, for his invaluable guidance, constructive criticism, and advice in completion of the work.

I also acknowledge the Master's thesis of Anurag Bhatt, who worked on related topics and through which I came to understand Item Response Theory better.

I sincerely thank all the faculty and staff members of the IME department for their direct and indirect assistance during my M.Tech.

I would also like to thank my friends and batchmates with a special mention to Rhit, Deepak, Narayan, Manish, Sanket, Jitendra and Ankush for making my two years of stay at IIT Kanpur an incredibly enriching journey.

Most importantly, I would like to thank my parents and my younger sister for their continuous support and belief in me.

Finally, I would like to thank Lionel Messi, arguably the greatest footballer on earth, who provides me continuous inspiration to work hard and remain humble in life.

Arijit Ganguly

June 2019

Table of Contents

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

In India, there is widespread and growing usage of multiple-choice questions in examinations for assessment of candidates' ability. These examinations are very important because they serve as a gateway to admissions to prestigious colleges and universities for higher education. These examinations are also used as a recruitment tool by essential government agencies. Traditionally, these exams use formula scoring as a measure of candidates' ability. GATE (Graduate Aptitude Test in Engineering) is an examination which uses the formula scoring method.

In this thesis, I present my portion of the research work done in a project that was jointly executed with Rhit Sanyal. In this combined project, we have conducted a post-facto analysis of GATE data from 2015 to suggest an alternative assessment policy to formula scoring in GATE. In my portion of the work, I have taken the raw item response data, and studied the question paper, and conducted a detailed descriptive analysis of the data. Subsequently, I have implemented a Tree-based Item Response Model, which is a particular type of Item Response Theory (IRT) model. I fitted this model to GATE 2015 dataset of 15 disciplines using packages available in R and compared with other Item Response Theory models to find the best model representing the dataset using three measures of fit to assess the statistical fit of the models. Rhit Sanyal, subsequently used the best model to generate candidate ability estimates and proposed a composite score as an alternative to traditional formula scoring in GATE. This composite score is a combination of the different latent traits provided by the best model. Using this score, he developed a two- dimensional characterization of a candidate's performance, which can be potentially acceptable of the various stakeholders of the GATE.

1.1 Graduate Aptitude Test in Engineering

GATE (Graduate Aptitude Test in Engineering) is a graduate level examination that is administered yearly in India. It mainly tests the fundamental understanding of various subjects in the undergraduate curriculum in disciplines. The GATE formula score serves a dual purpose. It is used primarily for admissions to a variety of post-graduate programs (e.g., Master of Engineering, Master of Technology, Doctor of Philosophy) in several disciplines of Engineering and Sciences. The second purpose is that several Indian public sector undertakings (PSUs) consider the GATE score for recruitment of graduate engineers at entry-level positions in these organizations. Thus, GATE serves for both employment and education in India.

GATE is administered in a collective annual effort by seven of the Indian Institutes of Technology as well as the Indian Institute of Science in Bengaluru. All these institutions are supported directly by the Ministry of Human Resource Development of the Government of India. Stakeholders of the GATE include the candidates appearing for the exam, the institutes coordinating the effort, the professors who set the question papers in confidential meetings, the academic institutions that use GATE scores for admissions, and the companies that use GATE scores for hiring. Since GATE is a fully government-supported activity, it is open to public scrutiny. GATE questions, the possibility of security breaches, the policy regarding GATE score usage, etc., are all matters of broad public interest. GATE has even been discussed in the Indian Supreme Court in response to questions and complaints raised by various parties.

In 2015, GATE was conducted by Indian Institute of Technology Kanpur in 23 disciplines. All disciplines combined 927,580 candidates registered for the examination, and 804,463 students appeared. It is a three-hour long examination conducted for all appearing candidates simultaneously. Examination for disciplines with a large number of candidates is generally held in multiple sessions. However, the 15 subjects which we have considered for our analysis do not have multiple sessions. GATE 2015 exam had 65 questions for each subject, with a maximum possible score of 100 marks.

Of the 65 items, 35 carried two marks each and 30 carried one mark each. Some were multiple choice questions (MCQ), which had four options out of which only one was correct, and others were keyed-in numerical answer type (NAT)questions. The answer of NAT questions had to be typed in with a virtual keypad which was provided. Responses within a given range were considered as correct. The MCQ's had negative marking (minus one-third of the question’s marks). The NAT's did not have negative marking because the probability of guessing the right answer was minimal. For both types of questions (MCQ's and NAT's), there were omitted, attempted-incorrect, and attempted-correct answers in the data.

1.2 Objective of the Thesis

As we know that GATE uses traditional formula scoring method, which is based on Classical Test Theory. The CTT has certain theoretical limitations which will be discussed in the Literature review. Apart from this, formula score suffers from some practical limitations which must be taken into account. First, formula scoring does not take into account question difficulties and question discriminations. If two candidates get the same score in spite of one candidate attempting easy questions and another candidate attempting difficult questions, both are given the same priority.

Secondly, if GATE is administered to a large group of test takers (say, tens of thousands), and certain desirable outcomes are possible for only the very top performers (e.g., several hundred, i.e., the top 1-2 percent), then we must distinguish between individual candidates in the upper tail of the distribution of the formula score. There, after a tiny proportion of the top performers are selected, even a slight further lowering of cutoff scores may induct a relatively large number of candidates, amongst whom subsequent secondary selections must be made.

All the limitations mentioned above can be easily tackled by using Item Response Theory (IRT) models to the GATE dataset. The IRT models estimate latent traits, which are a measure of the candidate ability. These models use essential factors which affect candidates’ performance. These factors include question difficulties and question discrimination. Hence, question or item parameters can also be evaluated by IRT models. Thus, IRT can perform not only candidate evaluation but also item evaluation. Also, the ability estimates generated from IRT models can also be used as a secondary score to distinguish between candidates having GATE scores with a negligible difference. Standard IRT models determine single ability estimate per candidate, but in reality, multiple factors might be influencing the performance of a candidate. Hence, we will propose a linear Tree-based IRT model having a two-dimensional ability estimate per candidate. This model will follow a tree structure which represents the decision making process in attempting a question. We will be comparing our proposed model with traditional IRT models. We will also analyze distinct varieties of Tree-based models. Finally, we will be selecting a model that fits the data most accurately and consistently for all the GATE disciplines considered.

GATE questions are set afresh every year. No questions are repeated, at least in principle. GATE questions, therefore, cannot be pre-tested on a separate population. Question parameters like difficulty must be estimated from the same dataset as the actual exam. However, GATE datasets are large, which makes parameter estimates more reliable. In the subjects included in the present study, every test taker was given the same set of questions. In some other subjects with many more test takers, two or three sets of questions were used. Those subjects are not included in this preliminary study, for simplicity. We have performed the *post-facto* analysis of response dataset of 15 disciplines which had a total of 66,084 candidates.

1.3 Data Description

In this section, we present a descriptive analysis of the item response data from the GATE 2015 examination. This part was jointly conducted by Rhit Sanyal and me. It has been reported for the completeness of the thesis.

The number of candidates appearing annually for GATE is indeed large: it was slightly above 800,000 for the all 23 GATE papers in the year 2015. Our analysis is on 15 subjects which had a smaller number of candidates. We left out large datasets like Mechanical Engineering (ME), Electrical Engineering (EE), Computer Science Engineering (CS), and a few others because our system configuration had limited processing speed. The total number of candidates who appeared for these 15 subjects is 66,084.

We have examined the GATE papers for the 15 subjects listed in Table 1. The smallest sample was $n = 527$ for EY (Ecology and Evolution), while the largest sample was $n = 16217$ for IN (Instrumentation Engineering). We note that the highest mean score is for MT (Metallurgical Engineering) at 39.39/100, while the lowest is for AG (Agricultural Engineering) at 12.88/100. Thus, the numbers of candidates vary greatly across subjects, and there are large differences between the marks distributions as well.

In three of the subjects shown in Table 1, there were multiple optional sections. Geology and Geophysics (GG) had one mandatory section and two optional sections out of which only one had to be attempted. Engineering Sciences (XE) had one mandatory section and six optional sections out of which two had to be attempted. Life Sciences (XL) had one mandatory and five optional sections out of which two had to be attempted. In our analysis, for each subject, we have selected the one combination which had the maximum number of test-takers. This process was followed for these three subjects.

Table 1: Subject wise details of GATE 2015 dataset.

Code	Subject name	Total number	Mean score	Median	St. Dev.
AE	Aerospace Engineering	3970	14.32	11.67	12.24
AG	Agricultural Engineering	1196	12.88	10.00	12.06
AR	Architecture and Planning	4147	32.60	32.33	12.58
BT	Biotechnology	8923	15.77	14.67	10.41
EY	Ecology and Evolution	527	23.78	22.00	14.89
GG	Geology and Geophysics (Geology)	4917	18.25	15.67	13.68
IN	Instrumentation Engineering	16217	13.90	11.00	11.58
MA	Mathematics	4899	15.74	14.67	9.91
MN	Mining Engineering	2361	21.67	17.33	15.85
MT	Metallurgical Engineering	3745	39.39	38.00	20.10
PH	Physics	7684	16.28	15.00	10.96
PI	Production and Industrial Engg	3185	15.62	13.67	11.27
TF	Textile Engg and Fibre Science	1111	20.73	20.00	11.19
XE	Engineering Sciences (Material Science and Polymer Science)	1015	19.36	18.00	11.07
XL	Life Sciences (Biochemistry and Microbiology)	2187	15.71	14.33	11.06
Total		66084			

The number of NAT questions was lowest for Geology and highest for Mathematics (see Table 2). Each question had only one right answer, although the NAT questions sometimes allowed a numerical range. The MCQ’s had negative marking (minus one-third of the question’s marks). The NAT’s did not have negative marking because the probability of guessing the right answer was very small. For both types of questions (MCQ’s and NAT’s), there were omitted, attempted-incorrect and attempted-correct answers in the data. In the IRTree and IRT modeling below, we have treated the MCQs and NATs on a uniform footing, for reasons that we will explain in due course. The scores that were obtained using the mentioned marking scheme are referred to as “formula score” in the thesis.

Table 2: Numbers of MCQ and NAT questions for each subject.

Subject Code	No. of MCQ	No. of NAT
AE	36	29
AG	37	28
AR	48	17
BT	48	17
EY	52	13
IN	40	25
MA	30	35
MN	37	28
MT	45	20
PH	42	23
PI	43	22
TF	46	19
GG	57	8
XE	45	20
XL	51	14

Histograms of formula scores in all subjects are shown in Figure 1. The choice of numbers of bins is based on accommodating all the data within a reasonable vertical scale. The number of bins shows a large variation since the number of candidates varies a lot across subjects (from 527 for EY to 16217 for IN). The shapes of the histograms may be of interest to a large section of GATE stakeholders.

Cumulative distributions of formula score obtained for all subjects are shown in Figure 2. In this figure, the scales are uniform on the horizontal and vertical axis, and a more direct comparison of the marks distribution across subjects can be performed.

Next, we have analyzed the relation between the number of questions attempted and the formula score using Three-Dimensional Histogram in Figure 3.

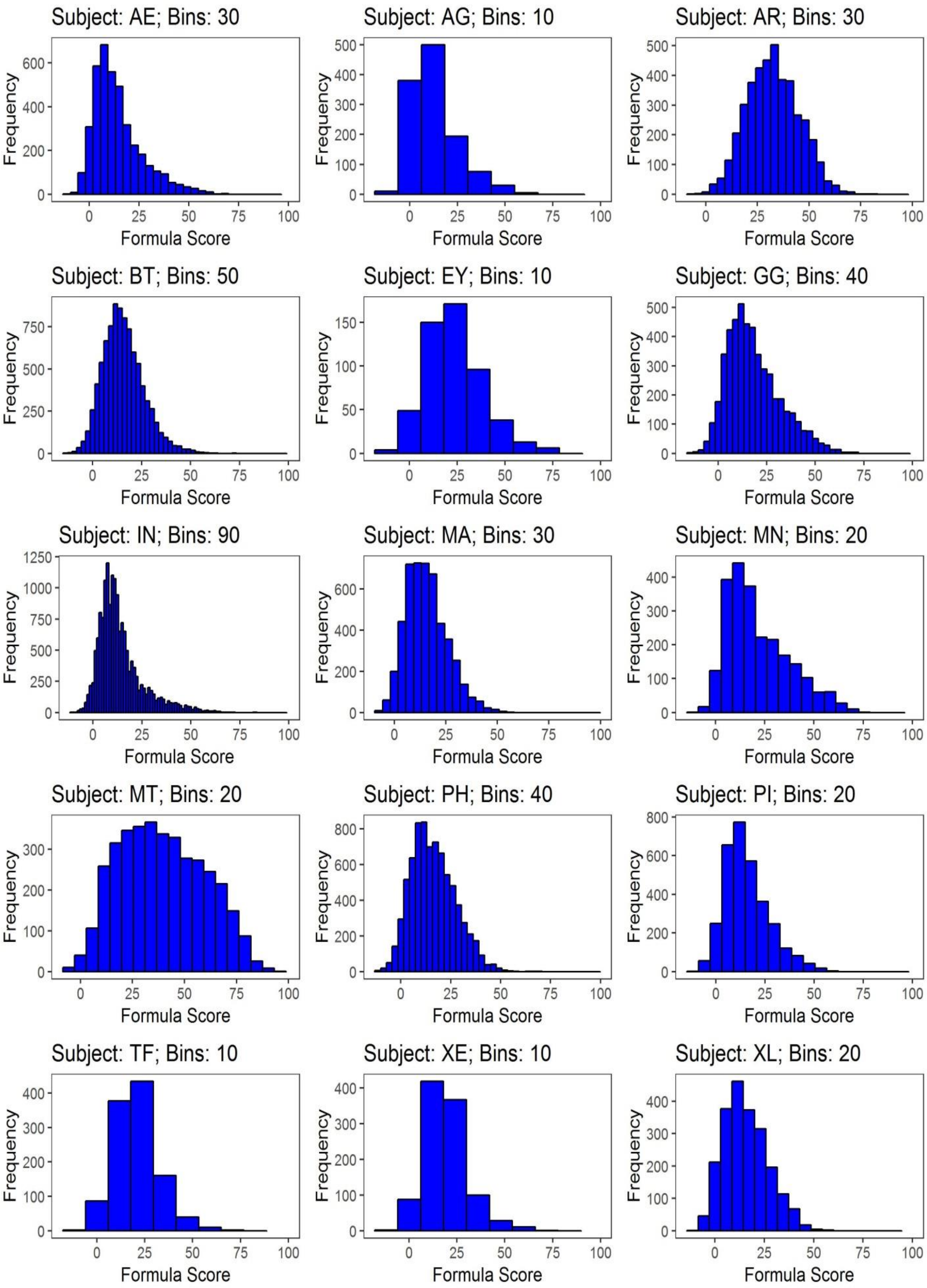


Figure 1: Histograms of formula scores for 15 subjects.

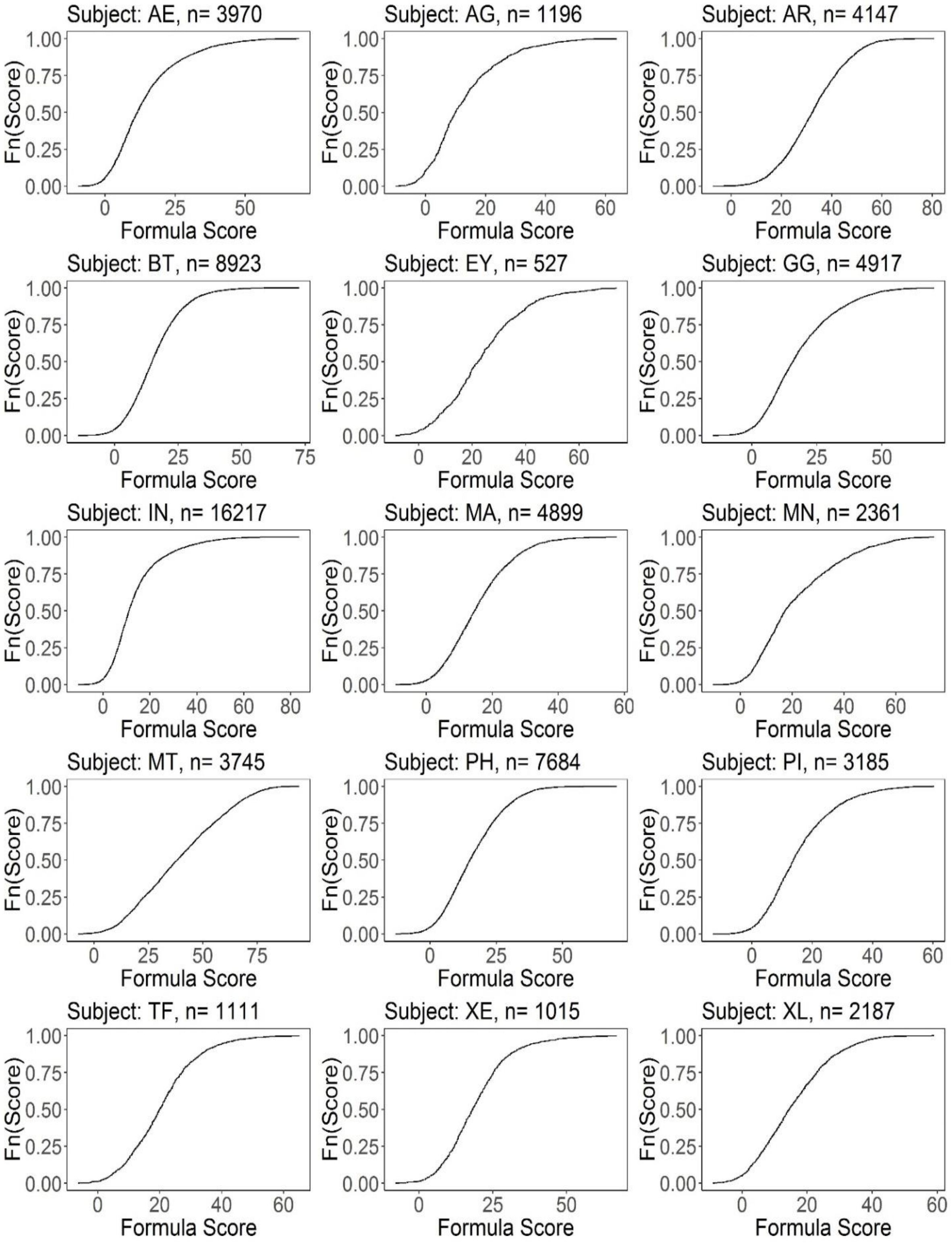


Figure 2 : Empirical cumulative distributions of formula scores for 15 subjects.

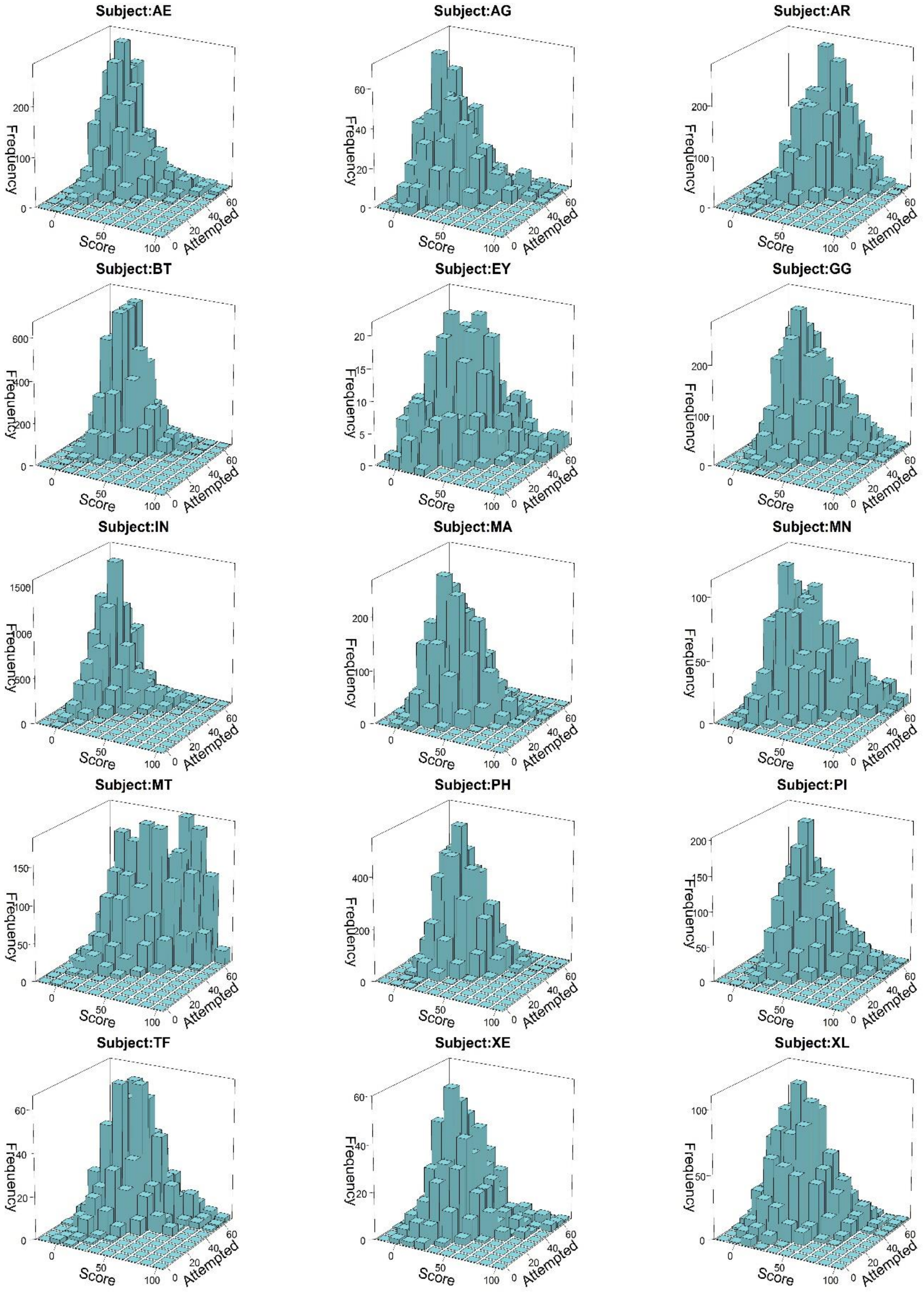


Figure 3: Frequencies of questions attempted against formula scores for all 15 subjects.

Figure 3 shows there are many candidates who attempt large numbers of questions and yet get low final scores. Relatively fewer candidates achieve high scores; some of them attempt relatively fewer questions, and presumably avoid losses due to negative marking. We can infer that, since there are several questions and relatively little time, candidates may adopt different strategies for scoring well. One extreme is to guess often, possibly after eliminating one or more options. Relatively weaker students might guess on more questions, especially if they expect their "true" score to be too poor to be useful. Conversely, strong candidates may avoid guessing to limit negative marking penalties.

So, it is clear that since GATE is a speeded test, with several questions and relatively little time, candidates can use different strategies based on their strength and weakness. One of the strategies can be guessing of MCQ options through the process of elimination. Relatively weaker students might use this strategy more often compared to candidates with stronger proficiency. During the exam, a candidate can switch back and forth between all the questions. So one of the strategies could be attempt questions on more familiar topics and then turn back to questions on less familiar topics. We have no information about the strategies of individual candidates. Also, we do not know the reason behind omitting a particular question by a specific candidate.

For these reasons, the IRTree modeling approach with two latent traits or propensities seems intuitively suitable for the GATE, because the first propensity determines the probability that the question will be attempted and the second propensity determines the probability of the attempted question being correctly answered. The first propensity models some combination of the candidate's strategy as well as speed; and the second one reflects actual proficiency, academic ability, and accuracy. Since both traits are important for good performance on the exam, we expect the IRTree model to fit the data well.

CHAPTER 2

LITERATURE REVIEW

In this chapter, we have presented a discussion of papers related to educational evaluation. Also, we have focused on papers related to Tree based Item Response Theory (IRTree) models. The entire section has been divided into sub-parts. At first, we talk about different ways to analyze multiple choice question assessments, including the two most widely used frameworks- the Classical Test Theory (CTT) and the Item Response Theory (IRT). Then we discuss the development and application of Item Response Theory. We discuss the research papers related to different IRTree models- their development and implementation. Finally, we look into the papers dealing with multiple ways to deal with missing responses in IRT models.

2.1 Theoretical Frameworks of Educational Evaluation

Ding and Beichner (2009) in their work, reported the five approaches of analyzing educational evaluation involving multiple choice test data in large scale assessment examination. They are as follows-

- Classical Test Theory (CTT) – According to this theory, the total test score is composed of a true score and a random component. The main objective of this is to analyze is an assessment is reliable and discriminating.
- Factor Analysis –This statistical technique is mainly used as an interpretation tool when there is a large number of items. So this reduces them to lower several common factors which are a linear combination of highly correlated items explaining most of the covariance between items.
- Cluster Analysis – This technique is used when the objective is to group examinees into different groups with distinguishable characteristics.

- Item Response Theory (IRT) – Most widely used statistical technique which is used to measure item parameters and examinees' latent ability using various models, mainly logistic models.
- Model Analysis–It measures the examinees' ability to answer isomorphic questions. It not only analyses attempting correct or incorrect but also the different approaches used by an examinee to solve the question.

Fan (1998) performed an empirical comparison between the Item Response Theory (IRT) and the Classical Test Theory (CTT) based on item and person statistics. He showed that the two frameworks are highly similar in terms of comparability and degree of invariance. Although the CTT is used widely in ability assessment, the use of IRT has been increasing in recent years. The CTT suffers from some significant limitations- the candidate ability (total score) is dependent on the sample, and the item statistics (difficulty and discrimination) are also reliant on the sample. While on the other hand, IRT has a better theoretical structure. It models the probability of a candidate answering an item correctly/incorrectly. There are many variations of IRT models available which can be applied in different unique situations- for example; dichotomous item responses can be modeled using one-parameter, two-parameter, and three-parameter model, while polytomous responses are modeled with partial credit and generalized partial credit model. IRT models overcome the major drawback of CTT models- the item statistics and person statistics are independent of the sample. This invariance property makes IRT models able to handle complex situations, which cannot be solved by CTT.

Champlain (2010), in his work, gave an overview of the CTT and the IRT by applying the two model frameworks on medical education assessment. He proved that with smaller sample sizes where the only purpose of the evaluation is to rank order candidates, CTT might be a better option compared to IRT. But in more complex situations with larger sample sizes, IRT is a much better option compared to CTT. This work revealed that both frameworks are complementary, and each presents various useful information.

2.2 Item Response Theory

Rasch (1960) developed the one-parameter logistic model (1PL), popularly known as the Rasch model which had a single item parameter called item difficulty and the candidate's latent ability which is the proficiency to be measured. This model assumes that item difficulty is the sole influencing factor on the candidate's performance.

Birnbaum (1968) extended the 1PL model by adding an item discrimination parameter and called it the two-parameter logistic model (2PL). But the 2PL model does not assume guessing behavior, which is evident in MCQ assessment. So he formulated the three-parameter logistic model (3PL) where the additional parameter called the pseudo-chance parameter, represents the guessing probability of answering correctly.

Masters (1982) extended the Rasch Model for dichotomous items for polytomous cases. This polytomous Rasch model was called the Partial Credit Model (PCM). In this model, the discrimination (slope) parameter had a fixed value constant across all items.

Muraki (1992) incorporated the varying discrimination (slope) parameter into the partial credit model and called it the Generalized Partial Credit Model (GPCM) which estimates parameters using Marginal Maximum Likelihood estimation with the EM algorithm. Based on test on simulated and real data, he showed that the GPCM model provides a better fit to polytomous item responses compared to PCM models.

Loken and Rulison (2010) formulated a 4 Parameter Logistic Model and estimated parameters in a Bayesian technique. They assumed the fourth parameter instead of considering the upper asymptote of the 3PL model as 1. Using simulated data, they showed that the 4PL model provides a better fit compared to 3PL and 2PL models and is more insightful compared to the other IRT models.

All the above works are regarding the unidimensional model, where only a single proficiency is measured. But in real life, assessments can measure multiple underlying latent abilities. Reckase (1997), in his work, proposed a Multidimensional IRT model intending to measure more than a unique ability of a candidate. He formulated a multidimensional version of the 3PL model in vector form. He applied this model on simulated data and found a better fit compared to unidimensional models. Also, this model was more insightful compared to unidimensional models.

Bolt *et al.* (2012), in their work, proposed a multidimensional version of the 2 Parameter Nested Logit Model and compared it with the nominal response model. The parameters were estimated using a Bayesian approach, and the model was applied to real life data.

2.3 Tree-Based Item Response Theory (IRTree)

De Boeck and Partchev (2012) formulated item response models having a tree structure which represents the sequential decision process of a candidate while answering a question. They developed the 1PL model having a linear response and nested response tree structure. Both unidimensional and multidimensional models were developed. These are generalized linear mixed models. Thus they can be solved using GLMM packages available in R- lme4. They also developed an R package –'*irtrees'* to implement the item response tree structure. But this has two main limitations- First is the ordinal response categories use a sequential approach but not the partial credit approach, which is used in GPCM. And second, 2PL and 3PL models cannot be implemented.

Böckenholt (2013) proposed a similar model to De Boeck and Partchev (2012) for modeling polytomous item responses using a linear or nested tree structure. The main difference was that Böckenholt used one parameter probit model, while De Boeck used a one-parameter logistic model. Also, he used Mplus software for estimation.

Bates *et al.* (2014) described the formulation and representation of linear mixed models. The mixed model means that the model uses both fixed effect and random effects in the expression. They also explained about the R package '*lme4'* which is used to fit and analyze linear and non-linear mixed models. This package is well documented, efficient, and easy to use for analyzing mixed models.

Okumura (2014), in his empirical study, applied De Boeck's IRTree model on PISA 2009 Japanese data to study omission tendency and reading ability. The model was solved using the GLMM techniques from the lme4 package in R. The results showed that students were more likely to omit open-ended questions than close-ended questions although open-ended questions had fewer item difficulties compared to close-ended questions.

Jeon and Rijmen (2016) developed a new R package –'*flirt'* for flexible item response theory. This package overcomes the limitations of '*irtrees*.' It implements both generalized linear and non-linear mixed models. Thus 2PL, 3PL, and partial credit models can be used. Also, it uses a more efficient maximum likelihood estimation, which reduces computation time.

Jeon and De Boeck (2016) developed a generalized version of IRTree model and implemented it using the R package – '*flirt'*. This structure is extremely flexible and is node specific; that is, it allows as many nodes as required, a different IRT model at each node, the multidimensional or unidimensional latent variable at individual nodes or shared across nodes, similarly for item parameters. They also developed a bifactor structure in which there is a common latent variable as well as node-specific latent variables. It uses a modern expectation- maximization algorithm (EM), which makes computation much faster. However, in our analysis, we stick to IRTree 1PL model developed by De Boeck et al. (2012). The generalized form can be explored in future work.

Jeon, De Boeck, and Linden (2017) worked on an application of the generalized item response models. They proposed a statistical method to model the answer change process of examinees using the tree structure of generalized IRTrees. The results proved that the decision at each node is the result of a unique decision process. This approach can also be used to model answer change behavior in psychological and behavioral assessment in general.

Böckenholt and Meiser (2017) applied the tree based item response models on Likert type items which had five outcomes possible on a scale of 1 to 5. They used the tree structure to decompose the responses into a sequential decision process. Different types of tree structures were considered. Finally, the best fit structure was selected.

2.4 Modeling Missing Responses

Omitted responses in examinations are frequent. If they are non-ignorable, then removing them from the IRT models can have serious effects. Holman and Glas (2005) implemented IRT models for handling non-ignorable missing responses. They used Partial Credit and Generalized Partial Credit Models for modeling the missing and observed responses. In simulation studies, they showed that ignoring MNAR included biases in the estimation of item parameters. The biases increased as a function of the correlation between the latent trait to be measured and the latent trait associated with omission. The missing data models increased the model fit to a considerable extent.

Glas and Pimentel (2008) worked on speeded tests that are tested with time limits and showed that nonignorable missing responses are present, which if ignored, can lead to biased estimates. They modeled the observed and absent responses with separate IRT models and proved that biases could be significantly reduced. For modeling the missing-ness, they used a sequential model with linear restrictions. This was performed on both simulated and real-world data.

Köhler, Pohl, and Carstensen (2015) in their work, tested whether the missing propensity is unidimensional and whether the missing propensity and ability are bivariate normally distributed. The results showed that although the unidimensionality assumption of missing propensity was not valid for all domains, it hardly affected ability parameter estimates. Thus considering missing propensity as unidimensional is justified. It also showed that the bivariate assumption was violated; hence, an appropriate distribution should be assumed to the missing propensity in the model.

Debeer, Janssen and De Boeck (2017) showed that if there are responses which are missing not at random (MNAR) are modeled assuming missing at random (MAR), that can result in bias estimates and poor model fit. So they proposed four IRTree models which considered two types of omissions- skipped items and not reached items. Two of them - Item selection model (ISM) and Continuing effort model (CEM) were applied on PISA 2009 dataset. They showed that the ISM model was the best representation when omissions are MNAR.

CHAPTER 3

PROBLEM DEFINITION AND METHODOLOGY

As the literature review shows that there is a very large amount of work in the application of Item Response Theory to exam datasets and relatively less work on applications of IRTree models to exam datasets. In any case, there is very little application of the IRT or IRTree models to Indian datasets. Since the IRTree models are more intuitively pleasing, we are going to focus on the application of such models to the GATE data.

3.1 Problem Definition

Initially, we have used the GATE item response data of the 15 subjects and converted it into polytomous response dataset. Then we have fitted the Partial Credit Model (PCM) and Generalized Partial Credit Model (GPCM) using the *TAM* package available in R, which is developed by Robitzsch *et al.* (2019). Further, we have fitted variations of linear Tree-Based Item Response Models (IRTree) using the *'irtrees'* package of R developed by De Boeck (2012). In IRTree modeling, first, we have implemented four different unidimensional and multidimensional models proposed by De Boeck (2012). After selecting the best out of the four models, we have tried different possible linear tree structures of the same model. Finally, we have chosen the best tree structure. We have compared the different types of models using model fit statistics – Akaike Information Criteria (AIC), Bayes Information Criteria (BIC) and Log-Likelihood statistics to come up with the best model that describes the data most appropriately. Using the best model, we have estimated the ability of a candidate or the latent traits in GATE assessment.

3.2 Methodology

In this section, we have explained the pre-processing of the data and implementation of all the models that were applied to GATE 2015 datasets for 15 subjects.

3.2.1 Data Pre-processing

Before applying the models, we have converted the GATE response data into a polytomous form. Here, each column represents each question or item. So there are 65 columns in every dataset which represents 65 questions. Each row represents a response to the questions by an individual candidate. So, the total number of rows is equal to the total number of test-takers. There are three possible responses to each item by a candidate. They are attempted and correct, attempted and incorrect and omitted. We have coded this response in the following way-

Table 3: Polytomous response format

Attempted Correct	Omitted	Attempted Incorrect
2	1	0

This type of polytomous response format represents the ordinal response category. It was also used by De Boeck *et al.* (2012) in their IRTree model formulation. It is more intuitive because it explains the ordering of the three categories. Attempting correctly results in a positive score. Hence, it has the highest weight. Omission awards zero score. So, it has a lower weight. Finally, attempting incorrectly results in a negative score. So, it has the least weight. The polytomous response coding could have been done in different order also, which we have explored later. But in our model, we are sticking to this format.

It is known that omission can occur because of two inherent reasons- skipping a question and not reaching a question due to time shortage. Thus, missing propensity can be assumed to be multidimensional representing two different latent traits. But this is not necessary. Köhler *et al.* (2015) showed that although the unidimensionality assumption of missing propensity was not valid for all domains, it hardly affected ability parameter estimates. Thus, considering missing propensity as unidimensional is justified.

In the data, each item had three possible outcomes: omitted, attempted-incorrect, and attempted-correct. In GATE's formula scoring, negative marking is present for MCQs (multiple choice questions) but not for NATs (numerical answer types). This is because the probability of randomly guessing the correct answer in MCQs is 25%, while the probability of guessing the correct answer for NATs is negligibly small. Prompted by the negative marking scheme, one might initially think that the MCQs should be modeled with three outcomes while the NATs should be modeled with two results (with an omitted answer treated as incorrect). However, negative marking is not the only penalty for attempting a question. For NATs, randomly guessing is wastage of time, and most candidates might avoid it. We have assumed that a candidate attempting a NAT question and getting an incorrect answer had invested some effort and time, which should be accounted for. For these reasons, we have decided not to be influenced by the formula scoring scheme, and to treat both MCQs and NATs on equal grounds, for both IRTree and IRT (PCM and GPCM) modeling.

3.2.2 Partial Credit Model (PCM) and Generalized Partial Credit Model (GPCM)

The Partial Credit Model (PCM) was formulated by Masters (1982). It is a generalization of the Rasch 1PL Model, which was proposed by Rasch (1960). The Rasch model can only be applied on dichotomous items ("correct" vs. "incorrect). But the PCM can be extended to polytomous items such as GATE response. Hence, we have considered the PCM as our first model and applied on the polytomous GATE data for all the 15 subjects.

At first, we have discussed the Item Response Theory (IRT) models which have been well explained by DeMars (2010) in his book on Item Response Theory. The most basic of them is the one parameter logistic model, which is also called the Rasch 1PL model. The probability that the j^{th} candidate's answer to the i^{th} question is correct would be taken as

$$P_{ij} = \frac{1}{1 + e^{-z_{ij}}} \;, i = 1,2, \dots \dots \dots p. \qquad j = 1,2, \dots \dots \dots N. \tag{1}$$

Where $z_{ij} = \theta_j - b_i$, θ_j represents the latent trait or ability of candidate j, which is to be measured, b_i represents item difficulty of item i. It is the measure of the difficulty of the question. The more difficult the question, the higher will be the value of b_i, p is the total number of items, and N is the total number of candidates or test-takers. P_{ij} is represented using a sigmoid curve with values between 0 to 1 on the ability scale. When $\theta_j \gg b_i$ then the probability of getting item i correct is almost close to 1 and vice-versa. This curve is also known as the Item Characteristic Curve (ICC). According to this model, the item difficulty is the sole item characteristic that affects the performance of a candidate.

This model was modified into the two-parameter model (2PL) by Birnbaum (1968), which has an additional item characteristic called item discrimination. This parameter represents the slope of the ICC. Questions with higher discrimination have a steeper slope compared to questions with lower discrimination. The transition from low to high probability occurs over a smaller range of θ_j if the discrimination a_i is higher. The 2PL model can be given as-

$$P_{ij} = \frac{1}{1 + e^{-z_{ij}}} \quad , i = 1,2, \dots \dots \dots p. \qquad j = 1,2, \dots \dots \dots N. \tag{2}$$

Where $z_{ij} = a_i(\theta_j - b_i)$. Here the discrimination parameter represents the slope of the ICC. An important feature of these models is the *property of invariance*. This means that item parameter estimates are independent of the distribution of candidate ability in the sample. Also, the ability estimates are independent of the test items.

There are two main assumptions of the above IRT models. They are as follows-

- Unidimensionality- These IRT models can measure only a single latent trait or ability. Multiple abilities cannot be measured. But this is practically not possible because the performance of a candidate is influenced by various factors like cognitive skills, the tendency of guessing, anxiety, and many others. That is why Multidimensional models give much better results which we will see in the case of Multidimensional IRTree models.

- Local Independence- This means that the set of abilities taken into account in a specified model are the only factors that influence candidates' response to items. So if unidimensionality assumption holds, there is only one ability in the entire latent space.

Estimation of ability and item parameters is the next step after fitting the model (Hambleton *et al.*, 1991). There are several ways of item parameter estimation which are as follows-

- Joint maximum likelihood estimation procedure – Proposed by Lord (1974). The item and ability parameters are estimated simultaneously. This is applicable for both 1PL and 2PL models.
- Marginal maximum likelihood estimation procedure– This was proposed by Bock and Atkin (1981). Initial distribution is assumed for ability estimates. With this, the item parameters are determined. Eventually, abilities are estimated with the item parameters. This procedure is also applicable for 1PL and 2PL models.
- Bayesian estimation procedure – Proposed by Mislevy (1986) and Swaminathan & Gifford (1982). It overcomes the problems faced by other methods like incorrect estimation or non-convergence. This procedure is also applicable for 1PL and 2PL models.

3.2.2.1 Model Formulation

Although 1PL and 2PL models can be applied on dichotomous data, they are not designed for polytomous data such as GATE data which has three ordered response categories. Hence, we have used the PCM and GPCM, which are polytomous extensions of the 1PL and 2PL models, respectively (Nering *et al.*, 2011). If there *are m* number of response categories, then according to PCM the probability that a candidate with ability θ responding in category k is given as-

$$P_{ik}(\theta_j) = \frac{e^{\sum_{j=0}^{k}(\theta - b_{ik})}}{\sum_{i=0}^{m-1} e^{\sum_{j=0}^{i}(\theta - b_{ik})}} \qquad (3)$$

Where b_{ik} is the difficulty parameter for category boundary parameter k of item i.

The GPCM is a further modification of the PCM where the additional discrimination (slope) parameter is incorporated into the existing PCM. The GPCM assumes that the probability that a candidate j with ability θ_j responding in category k is given as-

$$P_{ik}(\theta_j) = \frac{e^{\sum_{j=0}^{k} a_i(\theta - b_{ik})}}{\sum_{i=0}^{m-1} e^{\sum_{j=0}^{i} a_i(\theta - b_{ik})}} \tag{4}$$

Where a_i is the discriminating parameter of item i.

3.2.2.2 Model Implementation –

The PCM and GPCM had been applied on the polytomous datasets for the 15 GATE papers. For this purpose, we have used the IRT package 'Test Analysis Module' (TAM) (Version- 2.12-18) which is available in the free statistical programming language R. This package was developed by Robitzsch *et al.* (2019). Both unidimensional and multidimensional models can be applied to this package. 1PL and 2PL models can be implemented for dichotomous data, while PCM and GPCM can be implemented for polytomous data. This package uses Joint maximum likelihood estimation and Marginal maximum likelihood estimation techniques. We have used the Marginal maximum likelihood technique (MML) for estimation of item parameters. Then we have used Expected a-posteriori (EAP) technique for estimation of candidate ability or latent traits.

The main steps used for model fitting are-

- For Partial Credit Model – The PCM model was fitted to the data using the **tam.mml** function, where the parameter '*irtmodel'* was specified as "1PL". This signifies that only one difficulty parameter was estimated per item. Rest of the parameters were set to their default values.

- For Generalized Partial Credit Model – The GPCM model was fitted to the data using the function **tam.mml.2pl**, where the parameter '*irtmodel'* was specified as "GPCM". Rest of the parameters were set to their default values.
- TAM ability estimation - After building the model, the next step was to estimate candidate ability using **IRT.factor.scores** function. Here, the **model** is the already defined object which is created in the first stage. The option type gives three ways in which the ability estimation can be done. These are **MLE**: Maximum Likelihood Estimation, **EAP**: Expected A - Posteriori, and **WLE**: Weighted Likelihood Estimation. We use the option **EAP** in our analysis. This code has been repeated for both the PCM and GPCM models. The entire code can be found in Appendix A.

3.2.3 Tree-Based IRT Models (IRTree)

So far, we have explored only unidimensional models which can measure a single latent trait or candidate ability. But in reality, there is a vast number of factors which can govern the performance of a candidate in an examination. Also, we assume that correctly attempting a question is a two-stage decision process. First, is the decision to attempt or omit the question which can be governed by factors like usage of time, strategy, speed, and other factors. After attempting a question getting it correct requires proficiency, solving ability, calculation skills, and others. So the use of a two-dimensional sequential model is justified in this case. De Boeck *et al.* (2012) formulated IRTree structure for modeling an item response. They also implemented an efficient technique to solve the model and estimate item and ability parameters.

We have implemented the Linear response tree model. Response tree models, in general, assume that the response process is a sequential affair of going along the branches of a tree and finally reaching its end nodes. Linear response tree means that from an internal node, one branch leads to an end node, while the other one moves on to the following internal node. The outcomes of the left and right branch from a node are modeled using a one-parameter logistic function. This 1PL model is applied at each node. The latent traits θ are called *propensities* and may be vector valued, e.g., pairs of traits (propensities) θ_1 and θ_2 may be estimated for each candidate. We have used $\theta_{j,1}$and $\theta_{j,2}$for candidate j. Here, θ_j ($\theta_{j,1}$, $\theta_{j,2}$) follows a multivariate normal distribution. And, $\theta_{j,1}$ measures the *omission propensity* at the starting node, while $\theta_{j,2}$ measures the latent trait or ability at the first internal node.

The first decision, i.e., the decision to attempt at Node 1, occurs with a probability

$$P_{ij}^{(1)} = \frac{1}{1 + e^{-s_{ij,1}}}\ , \tag{5}$$

Where $s_{ij,1} = \theta_{j,1} + b_{i,1}$.

Subsequently, given that an attempt has been made, at Node 2, the probability that the answer is correct is taken to be

$$P_{ij}^{(2)} = \frac{1}{1 + e^{-s_{ij,2}}}\ , \tag{6}$$

Where $s_{ij,2} = \theta_{j,2} + b_{i,2}$.

The item difficulty parameters $b_{i,1}$ and $b_{i,2}$ are called *induction intensities*. Here $-b_{i,1}$ is interpreted as the intensity of omission induction while $b_{i,2}$ is interpreted as easiness (negative of difficulty). The sign convention of these parameters is opposite (positive) to that of conventional IRT models. This is merely a sign convention and has no consequences other than interpretation.

3.2.3.1 Model Formulation -

There are three ordinal response categories. 0 represents incorrect response, 1 represents omission, and 2 represents the correct response. Figure 4 represents our proposed linear IRTree structure. The first node - Node 1 is represented by ‘Question(Y_1^*)’. The top branch from Node 1 determines the probability of omission, while the bottom one leads to an internal node. So, the top branch is coded 0, and the bottom is coded 1. The internal node is Node 2, which is represented by ‘Attempted(Y_2^*)’ further leads to two branches. Upper one denotes incorrect response, while lower one denotes correct response. In the same way, upper is coded as 0 and lower as 1. Hence, omission (1) is coded as (0, NA). Incorrect response (0) is coded as (1,0) and correct response is coded as (1,1). The mapping matrix is shown in Table 4.

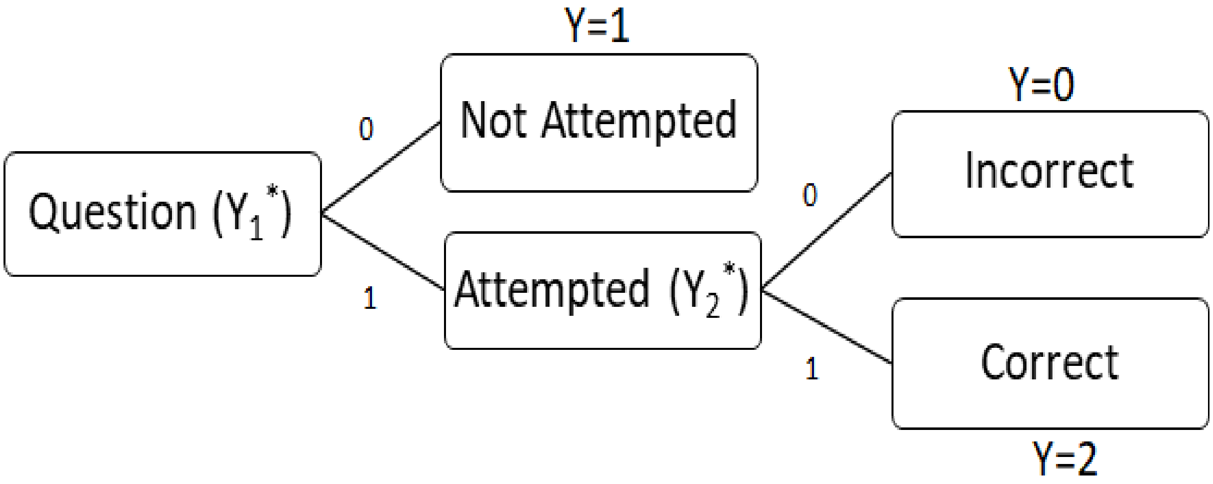


Figure 4: Linear IRTree structure for modeling of GATE data. Other structures are possible.

Table 4: Mapping Matrix for the above IRTree structure.

Response Outcomes	Node 1	Node 2
Y=1	0	NA
Y=0	1	0
Y=2	1	1

From this tree structure, we can easily define the probability of the three outcomes – omission, attempting correctly, and attempting incorrectly. The probability of omission of the *i^th^* question by candidate *j* is given by-

$$P_{ij}\left(Y_{ij} = 1 \,\middle|\, \Theta_j\right) = P\left(Y^*_{ij1} = 0 \,\middle|\, \Theta_{j1}\right) \tag{7}$$

The probability that candidate *j* attempts i^{th} question correctly is given by –

$$P_{ij}\left(Y_{ij} = 2 \,\middle|\, \Theta_j\right) = P\left(Y^*_{ij1} = 1 \,\middle|\, \Theta_{j1}\right) \times P\left(Y^*_{ij2} = 1 \,\middle|\, \Theta_{j2}\right) \tag{8}$$

Finally, the probability that candidate *j* attempts i^{th} question incorrectly is given by –

$$P_{ij}\left(Y_{ij} = 0 \,\middle|\, \Theta_j\right) = P\left(Y^*_{ij1} = 1 \,\middle|\, \Theta_{j1}\right) \times P\left(Y^*_{ij2} = 0 \,\middle|\, \Theta_{j2}\right) \tag{9}$$

After defining the IRTree structure, we have applied the proposed multidimensional linear IRTree model on item response of 15 GATE subjects. But we wanted to compare our IRTree model with other multidimensional and unidimensional IRTree models to find out the best one. So, we have also applied other IRTree models (De Boeck *et al.* 2012) and compared the results.

- Multidimensional Model – This is the proposed model which we have discussed above. It estimates two propensities at two nodes. Since this model has two induction intensities per question and two propensities per test taker, we have labeled this model as “2I,2P”.
- Unidimensional Model – Both the Multidimensional and Unidimensional model have the same tree structure. The only difference is that both the nodes in the unidimensional model are used to measure the same propensity. This model has been labeled “2I,1P”.
- Unidimensional Model with single intensity per item – This model is very similar to "2I,1P". The only difference is that the node- item interaction is eliminated. So, there is only a main fixed effect per tree. This model has been labeled as "1I,1P".

- Multidimensional Model with Random item effect – This model is similar to "2I,2P". The difference is that items are assumed to have random effects with a multivariate distribution. An additional term is introduced in the model equation, which accounts for the item variance per node. Also, an item correlation matrix is calculated for every model. This model has been labeled as "2IR,2P".

The implementation of the above models is shown in the next section. We have compared the four models based on model fit statistics (AIC, BIC, Log-Likelihood). The model having the optimum values of model fit statistics, has been selected. The entire code for the above procedure can be found in Appendix A.

Although the chosen modeling approach is expected to be the most suitable, there are two other IRTrees that could, in principle, be used to model the same data. After selecting the best model, we have tried out other potentially possible linear IRTree structures that could represent the polytomous response pattern. We have applied the best model chosen from the above process on two other possible IRTree structures and compared the fit statistics with our proposed one.

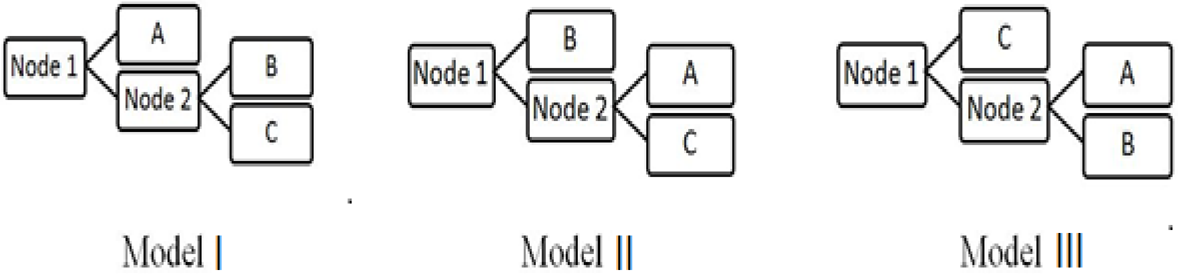


Figure 5: Three IRTrees for the same data, with outcomes A = omitted response, B = attempted-incorrect, and C = attempted-correct.

In Figure 5, the proposed IRTree model is on the left (Model I), while the other possible IRTree structures are Model II and Model III. As we know, in Model I, the first node accounts for the probability of attempting or omitting, and the second node accounts for the probability of attempting correctly or incorrectly. In Model II, the first node accounts for the probability of attempting incorrectly. In Model III, the first node accounts for the probability of attempting correctly.

Table 5: IRTree mapping matrices of the three structures

Model I			Model II			Model III		
Linear Response	Node 1	Node 2	Linear Response	Node 1	Node 2	Linear Response	Node 1	Node 2
Y=1	0	NA	Y=1	1	0	Y=1	1	0
Y=0	1	0	Y=0	0	NA	Y=0	1	1
Y=2	1	1	Y=2	1	1	Y=2	0	NA

We have already discussed the Mapping matrix of Model I, which is the proposed tree structure. In Model II, the top branch from Node 1 determines the probability of attempting incorrectly (Y=0). Hence, it is coded as (0, NA). Similarly, omitted response (Y=1) is coded as (1,0) and attempting correctly (Y=2) is coded as (1,1). In Model III, the top branch from Node 1 determines the probability of attempting correctly (Y=2). Hence, it has been given the code (0, NA). Similarly, omitted response (Y=1) is coded as (1,0) and attempting incorrectly is coded as (1,1). The mapping matrices of the three structures have been reported in Table 5.

These structures are fitted to the dataset using the same model which we have selected from the previous step. All three possibilities are compared based on model-fit statistics. The structure which represents the response pattern most accurately has been chosen. The R codes have been reported in Appendix A.

3.2.3.2 Model Implementation-

The implementation of the above IRTree structure has been performed using the package '*irtrees'* (Version- 0.1.0) which is freely available in R. The *dendrify* function under this package has been used. This function helps to convert the item response data matrix, which is in wide format to a long-form data frame where each row denotes a combination of a unique candidate, a unique question and a unique node. The IRTree model is a linear mixed effects model. So the Generalized Linear Mixed Model (GLMM) technique has been used for model fitting and parameter estimation. This has been implemented using the *'lme4'* package (Version- 1.1-21) in R, which is developed by Bates *et al.* (2015).

Maximum likelihood estimate technique has been used for parameter estimation. For the mixed effects model, the expression for maximum likelihood is an integral which is approximated using adaptive Gauss-Hermite quadrature. For numerical integration (Gauss-Hermite) within the routine, there is an integer parameter "nAGQ" that needs to be specified. Numbers greater than one result in more accurate output, but take colossal time. An input of nAGQ = 1 uses of the Laplace approximation. When nAGQ = 0, random effects are not integrated out. The random effects and the fixed effects are optimized in the penalized iteratively reweighted least squares (PIRLS) step, and estimation becomes quicker (see Bates *et al.* (2015), Stegmann *et al.* (2018)). For large data sets, the difference in run time is large, and the difference in estimated quantities is tiny. For example, for our second smallest data set corresponding to subject code XE (with 1015 students; see Table 1), the performance of nAGQ = 0 as compared to nAGQ = 1 is already remarkable. The former takes 1.3 minutes to run on an ordinary desktop PC while the latter takes almost 10.8 hours on the same machine; the AIC, BIC, and Log Likelihood values (model fit statistics) differ in the 6th, 6th, and 5th significant digits respectively; and the correlations between the person traits (propensities) obtained using the two algorithms are 0.9999996 for θ_1 and 0.9999958 for θ_2. There is thus no practical difference in results, and the computer run time is vastly different (almost a factor of 500). Accordingly, in our analysis using R, we used nAGQ = 0 for all subjects. All other parameters used in the study were default values of the R package.

In this section, we have presented presenting the implementation the four IRTree models, namely – "2I,2P", "2I,1P", "1I,1P" and "2IR,2P" and eventually the implementation of the three structures – Model I, Model II and Model III. As mentioned, we have used the *'lme4'* package in R. For the application of GLMM, we have used the **glmer** function. After fitting the model to the response datasets, we have estimated the latent traits and omission propensities of the candidates. The latent traits or propensities are modeled as random effects following a multivariate normal distribution. They have been estimated using Bayesian posterior mode estimate. We have used the **ranef** function for achieving this purpose. This function finds the modal estimates of candidate ability values. However, if one wishes, the variance estimates can also be easily computed using another function in 'lme4' package.

The four IRTree models are implemented at first. The GLMM equations of these models are given below.

- "2I,2P" Multidimensional Model – The modeling equation is given as

  Model1 <- glmer(value ~ 0 + item:node + (0 + node | person), + family = binomial(link="logit"), data = linrespT, nAGQ=0) (10)

  The term *item:node* denotes fixed effect at each node. For each paper, 130 (65×2) fixed effects have been estimated. The term *(0 + node | person)* is used to estimate the variance and covariance of the propensities. The term *binomial* suggests that the tree structure is of binomial type, that is it has only two outcomes (0 and 1). The term *link="logit"* is used to specify that the one-parameter logistic function is used at each node. The term *nAGQ=0* has already been explained.

- "2I,1P" Unidimensional Model – The modeling equation is given as

  Model2 <- glmer(value ~ 0 + item:node + (1 | person), + family = binomial(link="logit"), data = linrespT, nAGQ=0) (11)

  The term *(1 | person)* suggests that this model estimates a single latent trait per candidate.

- "1I,1P" Unidimensional Model with single intensity per item – The modeling equation is given as

  Model3 <- glmer(value ~ 0 + item + node + (1 | person), + family = binomial(link="logit"), data = linrespT, nAGQ=0) (12)

  This model estimates only a single fixed effect per item.

- "2IR,2P" Multidimensional Model with Random item effect – The modeling equation is given as

  Model4 <- glmer(value ~ 1 + (0 + node | item) + (0 + node | person), + family = binomial(link="logit"), data = linrespT, nAGQ=0) (13)

The term *(0 + node | item)* indicates that items are modeled with random effects. Thus item variance and item correlation matrix is also estimated for all the nodes.

All the four models, which are mentioned have been applied using the proposed IRTree structure, which is Model I. After selecting the best model, the other two structures – Model II and Model III were also tried out. The modeling equation remained the same. The only change that was made to represent the other structures was achieved by changing the mapping matrix, which has been shown in Table 5. This task was performed inside the *cbind* function in R. The rest of the process remains unchanged. Please refer to Appendix A for the R codes.

3.2.4 Measures of Fit of a Model

In this research, we have used the same measures of fit or Goodness-of-fit of all the models. These model-fit statistics are as follows-

- Akaike Information Criteria (AIC) – The AIC represents the amount of information that is lost when we are trying to fit a model that best represents the actual process. Hence lower the value, better is the model fit. It is given by –

  $$AIC = 2k - 2ln(L) \; ; \tag{14}$$

  Where, *k* is the number of estimable parameters in the model, and *L* is the maximized log-likelihood value. It is an optimization function which optimizes the number of estimable parameters. As the value of *k* increases, the first penalty term increases while the second term, which assesses the fit decreases. The AIC is developed to estimate the best accuracy among competing models or hypotheses. (Posada *et al.*, 2004).

- Bayesian Information Criteria (BIC) – The BIC is almost similar to AIC. It is an approximated value of log-marginal likelihood of a model. The model having the smallest value is the model having a maximum posterior probability, hence better. It is given by –

  $$BIC = kln(n) - 2ln(L) \; ; \tag{15}$$

Where *k* is the number of estimable parameters, *n* is the sample size, and *L* is the maximized log-likelihood value. (Posada *et al.*, 2004).

- Log-Likelihood – This function gives the logarithmic value of the probability that a particular model having a few estimable parameters is the best representation of the actual process, and it fits the data most accurately. Hence higher is the value; better is the model fit.

3.2.5 Data Analysis and Modeling Platform

In this research, we have used the freely available statistical tool R (version 3.5.2), and open-source software R-studio (version 1.1.463) for implementing R. All the packages we have employed in our research are freely available in the R repository. For the necessary processing of the data, we have used MS-Excel 2016.

CHAPTER 4

RESULTS - ANALYSIS AND DISCUSSION

In this chapter, we have explained the results of our analysis. Mainly we have discussed the model-fit statistics of each model and compared the models based on these measures of fit. Finally, we have selected the best model and used it to estimate candidate ability or latent trait. We have run all the models for all the 15 GATE subjects. A detailed description of the data can be found in chapter three.

4.1 Comparison of PCM, GPCM, and Multidimensional IRTree Models

As discussed in chapter three, we have executed all the required models and computed their model-fit statistics. At first, we have compared our proposed multidimensional (2I,2P) IRTree model with conventional unidimensional IRT models. These include PCM and GPCM. We found out that AIC values for the IRTree model were lower than that of the GPCM model. PCM had highest AIC values. This trend was seen for all the 15 subjects.

Similarly, BIC values were lowest for IRTree, followed by GPCM, and finally, PCM had maximum value for all the 15 subjects. Log-Likelihood followed a similar trend for 15 papers. It was maximum for IRTree, followed by GPCM and minimum for PCM. Also, we noticed that GPCM and PCM had very little difference in their fit statistics, but IRTree outperformed both the IRT models by a considerable margin. The results have been reported in Table 6. Based on this evidence, we have concluded that IRTree is a better model compared to PCM and GPCM for our GATE dataset. Hence, we have discarded these two models and continued with our proposed IRTree model.

Table 6: Comparison of model-fit statistics for PCM, GPCM, and Multidimensional (2I,2P) IRTree model

Subject	Model	AIC	BIC	Log-likelihood
AE	PCM	452775.9	453599.5	-226257.0
	GPCM	442057.6	443283.5	-220833.8
	IRTree	428299.0	429762.9	-214016.5
AG	PCM	141316.75	141983.11	-70527.37
	GPCM	138070.0	139062.0	-68840.02
	IRTree	132142.6	133437.8	-65938.3
AR	PCM	477878.3	478707.6	-238808.2
	GPCM	474478.1	475712.5	-237044.1
	IRTree	454772.5	456245.4	-227253.3
BT	PCM	1069198	1070120	-534468.9
	GPCM	1059621	1060998	-529616.6
	IRTree	1018529.9	1020097.8	-509131.9
EY	PCM	62689.21	63248.21	-31213.60
	GPCM	62193.76	63025.86	-30901.88
	IRTree	60180.4	61378.0	-29957.2
GG	PCM	597159.6	598011.2	-298448.8
	GPCM	591655.4	592923.0	-295632.7
	IRTree	570313.0	571803.8	-285023.5
IN	PCM	1842879	1843887	-921308.4
	GPCM	1811344	1812844	-905477.1
	IRTree	1720933.0	1722580.6	-860333.5
MA	PCM	547012.5	547863.6	-273375.2
	GPCM	537078.2	538345.1	-268344.1
	IRTree	514777.7	516256.7	-257255.9
MN	PCM	273612.8	274368.2	-136675.4
	GPCM	268832.5	269957.1	-134221.3
	IRTree	257878.4	259269.9	-128806.2
MT	PCM	424875.7	425691.5	-212306.8
	GPCM	419568.8	420783.3	-209589.4
	IRTree	404430.4	405893.0	-202082.2
PH	PCM	916505.0	917415.0	-458121.5
	GPCM	907661.9	909016.6	-453636.0
	IRTree	880063.5	881607.1	-439898.7
PI	PCM	367640.1	368434.8	-183689.1
	GPCM	361407.3	362590.2	-180508.7
	IRTree	345445.7	346877.9	-172589.8
TF	PCM	131068.78	131725.48	-65403.39
	GPCM	130334.13	131311.66	-64972.06
	IRTree	123691.2	124986.2	-61712.6
XL	PCM	257907.4	258652.8	-128822.7
	GPCM	256208.9	257318.5	-127909.5
	IRTree	247965.0	249346.0	-123849.5

XE	PCM	118089.70	118734.57	-58913.85
	GPCM	114757.35	115717.26	-57183.6
	IRTree	110978.9	112256.8	-55356.5

4.2 Comparison of Four Different IRTree Models

Next, we have compared the different linear IRTree models based on their fit statistics. The models in contention are – proposed “2I,2P” multidimensional model, “2I,1P” unidimensional model, “1I,1P” unidimensional model with single intensity per item and “2IR,2P” multidimensional model with random item effects. We noticed that both the multidimensional models had far better fit statistics compared to both the unidimensional models for all the subjects. Hence, we discarded the unidimensional models. Next, we observed that “2I,2P” model performed better based on AIC and Log-Likelihood criteria while “2IR,2P” model had better BIC values. This trend was consistent for the 15 subjects. The BIC values for “2IR,2P” were marginally better, but the computation time was three times higher on average than that of “2I,2P” model. Hence, we have selected the “2I,2P” multidimensional IRTree model as the most appropriate one. The model-fit statistics have been reported in Table 7.

Table 7: Comparison of model-fit statistics for four different IRTree models

Subject	Model	AIC	BIC	Log-likelihood
AE	2I,2P	428299	429762.9	-214016.5
	2I,1P	457620.4	459062.4	-228679.2
	1I,1P	477789.5	478527	-238827.8
	2IR,2P	429089.1	429166.1	-214537.5
AG	2I,2P	132142.6	133437.8	-65938.3
	2I,1P	141321.9	142597.5	-70529.9
	1I,1P	146073.8	146726.2	-72969.9
	2IR,2P	132742.3	132810.4	-66364.1
AR	2I,2P	454772.5	456245.4	-227253.3
	2I,1P	470375.8	471826.5	-235056.9
	1I,1P	499131.9	499873.9	-249499
	2IR,2P	455569.4	455646.9	-227777.7
BT	2I,2P	1018529.9	1020097.8	-509131.9
	2I,1P	1060844.8	1062389.2	-530291.4
	1I,1P	1119458.8	1120248.6	-559662.4
	2IR,2P	1019429	1019511.6	-509707.5
EY	2I,2P	60180.4	61378	-29957.2
	2I,1P	61788.6	62968.2	-30763.3
	1I,1P	64651.8	65255.1	-32258.9
	2IR,2P	60701.1	60764.1	-30343.5
IN	2I,2P	1720933	1722580.6	-860333.5
	2I,1P	1822468.2	1824091	-911103.1
	1I,1P	1913156.8	1913986.8	-956511.4
	2IR,2P	1721916.7	1722003.4	-860951.3
MA	2I,2P	514777.7	516256.7	-257255.9
	2I,1P	554390.1	555846.8	-277064
	1I,1P	575818.7	576563.7	-287842.4
	2IR,2P	515561	515638.9	-257773.5
MN	2I,2P	257878.4	259269.9	-128806.2
	2I,1P	274530.6	275901.1	-137134.3
	1I,1P	283830.9	284531.9	-141848.4
	2IR,2P	258597.1	258670.4	-129291.6
MT	2I,2P	404430.4	405893	-202082.2
	2I,1P	417748.3	419189	-208743.1
	1I,1P	431899.9	432636.7	-215882.9
	2IR,2P	405200.1	405277.1	-202593.1
PH	2I,2P	880063.5	881607.1	-439898.7
	2I,1P	918188.2	919708.7	-458963.1
	1I,1P	955980.8	956758.4	-477923.4
	2IR,2P	880919.7	881001	-440452.9
PI	2I,2P	345445.7	346877.9	-172589.8
	2I,1P	364300.5	365711.2	-182019.3
	1I,1P	383676.7	384398.2	-191771.3
	2IR,2P	346224.7	346300	-173105.3
TF	2I,2P	123691.2	124986.2	-61712.6
	2I,1P	128821.9	130097.4	-64279.9

	1I,1P	133821.3	134473.7	-66843.6
	2IR,2P	124319.7	124387.8	-62152.8
GG	2I,2P	570313	571803.8	-285023.5
	2I,1P	589677.7	591146.1	-294707.8
	1I,1P	615302.4	616053.4	-307584.2
	2IR,2P	571118.9	571197.3	-285552.4
XE	2I,2P	110978.9	112256.8	-55356.5
	2I,1P	117649.8	118908.5	-58693.9
	1I,1P	123286.6	123930.4	-61576.3
	2IR,2P	111595.9	111663.1	-55790.9
XL	2I,2P	247965	249346	-123849.5
	2I,1P	256033.6	257393.8	-127885.8
	1I,1P	267293.1	267988.8	-133579.5
	2IR,2P	248691.7	248764.4	-124338.8

4.3 Comparison of Three Possible IRTree Structures

After selecting the “2I,2P” IRTree model, we have compared the proposed linear IRTree structure with two other possible linear tree structures which could be used to represent the data. Since the proposed structure (Model I) describes the decision process of an item response most accurately, we know that it is the most suitable option. To assert our claim, we have run the "2I,2P" model for the other two structures and compared the fit statistics. As per our expectation, we found that Model I has outperformed both Model II and Model III based on all the three measures – AIC, BIC, and Log-Likelihood. Hence, we can claim that our proposed model provides the best fit to the GATE data for all the 15 subjects. The model-fit statistics have been reported in Table 8.

Therefore, from this comparative study of all the models, we have deduced that the “2I,2P” multidimensional linear IRTree model having the structure represented in Figure 4 is the best-suited model for the GATE dataset of all the 15 subjects. It has the best model-fit statistics and is also computationally faster compared to the other multidimensional IRTree model with random item effects. Thus, we have used this model to estimate item difficulties (*induction intensities*) for each question and estimate latent trait and *omission propensity* for each candidate. For AE GATE paper, we have reported the item difficulty values in Appendix B.

Table 8: Comparison of model-fit statistics for three possible 2I,2P IRTree structures

SUBJECT	MODEL	AIC	BIC	Log-likelihood
AE	I	428299	429762.9	-214016.5
	II	430921.6	432367.8	-215327.8
	III	429459.6	430926.2	-214596.8
AG	I	132142.6	133437.8	-65938.3
	II	133733	135027.1	-66733.5
	III	132852.2	134161.5	-66293.1
AR	I	454772.5	456245.4	-227253.3
	II	457323	458784.1	-228528.5
	III	455457.8	456918.6	-227595.9
BT	I	1018529.9	1020097.8	-509131.9
	II	1022162.5	1023722.6	-510948.3
	III	1019195	1020767.5	-509464.5
EY	I	60180.4	61378	-29957.2
	II	60520.3	61705.8	-30127.2
	III	60243.1	61431.7	-29988.5
GG	I	570313	571803.8	-285023.5
	II	574020.1	575502.6	-286877
	III	572239.1	573728.7	-285986.6
IN	I	1720933	1722580.6	-860333.5
	II	1730927.6	1732564	-865330.8
	III	1729512.9	1731167.5	-864623.5
MA	I	514777.7	516256.7	-257255.9
	II	515113.2	516598.4	-257423.6
	III	515694	517190.9	-257714
MN	I	257878.4	259269.9	-128806.2
	II	259602.9	260987.8	-129668.5
	III	258208.3	259602.2	-128971.2
MT	I	404430.4	405893	-202082.2
	II	406919.8	408369	-203326.9
	III	404655.4	406097.1	-202194.7
PH	I	880063.5	881607.1	-439898.7
	II	883400.3	884943.4	-441567.1
	III	883106.1	884660.2	-441420.1
PI	I	345445.7	346877.9	-172589.8
	II	347928.7	349349.1	-173831.4
	III	346655.1	348091.7	-173194.5
TF	I	123691.2	124986.2	-61712.6
	II	124423	125706.8	-62078.5
	III	124038	125328.6	-61886
XE	I	110978.9	112256.8	-55356.5
	II	111632.9	112906	-55683.5
	III	111224.6	112506.9	-55479.3
XL	I	247965	249346	-123849.5
	II	248741.4	250116	-124237.7
	III	248406.5	249790.5	-124070.2

CHAPTER 5

CONCLUSIONS

We have seen that the GATE formula score based on Classical Test Theory comes with certain practical and theoretical limitations. These limitations can be overcome by using Item Response Theory models. Hence, candidate ability based on latent traits estimated by IRT models can serve as an alternative to the current formula score. Thus, the paramount aim of this research work was to propose a robust and accurate statistical model that can be applied on GATE polytomous item response dataset. The model should be able to handle MCQ responses as well as NAT responses, give unbiased estimates of candidate latent trait, run on both small and large datasets, have less time complexity.

This analysis portrayed that the multidimensional linear IRTree with two propensities per candidate and two difficulties per item is not only statistically better, but also much more informative compared to PCM, GPCM, and unidimensional IRTree models. Here we can find two traits per candidate- *omission propensity* and latent trait. We can try to find the relation between these two traits to know whether they depend on each other. This latent trait can be used as a viable alternative to the conventional GATE score. Also, the *omission propensity* and latent trait can be combined to come up with an alternative ability measure to GATE score. The item difficulties are also beneficial. We obtain two item parameters – one is the intensity of omission induction, which governs the probability that a person will omit the question and the other is easiness (opposite of difficulty). These parameters can be used to evaluate the questions and study their effectiveness in candidates' ability assessment. Also, compared to the other multidimensional model with random item effects, our model is statistically better based on AIC and Log-Likelihood values and has much lower time complexity.

This model is consistent across all the datasets and is highly scalable. From the smallest dataset EY (527 candidates) to the largest dataset IN (16,217 candidates), the model consistently outperforms all the other models. Hence, it can be easily applied to much larger datasets like Mechanical Engineering, Electrical Engineering, and so on.

5.1 Limitations and Scope of Future Work

This research had to encounter a few limitations. They are as follows-

- We have used the response tree model developed by De Boeck *et al.* (2012), which uses a single difficulty parameter. We have discussed in the literature a generalized version of IRTree developed by Jeon and De Boeck (2015), which has a discrimination parameter, a bi-factor structure, choice of different functions in different nodes. But the only way to implement it is using R package *'flirt'* which is in the development stage and is unavailable. So, we have stuck to the 1PL version of IRTree. But in future more generalized version can be applied.
- We have applied our model only 15 GATE papers which were attempted by a relatively smaller number of candidates. IN had the maximum responses, which was 16,217. We have left out larger dataset like Mechanical Engineering (ME), Electrical Engineering (EE), Civil Engineering (CE) and few others which had a much larger number of candidates. This was because we had systems with limited memory and processing speed. Systems with better configurations can solve this issue.
- In the chapter of Methodology, we stated that one assumption for IRT models is local independence. But the test for local independence is quite a difficult and challenging task, and out of the scope of the thesis. Hence, we have assumed our models to be locally independent.

This work has tremendous future scope. First, the candidate ability estimates generated by the model can be used as an alternative to the traditional GATE score. If not as an alternative, this latent trait can be used as a secondary score to resolve ties between candidates with the same GATE score. A more generalized version of this IRTree model, like the 2PL version with a discrimination parameter or the bi-factor version, can be used when the required packages become available. The same model can be applied to much larger GATE datasets like ME, EE, etc. in future work. We have seen that our IRTree model performs consistently better over other models. Hence, this proposed model, with few modifications, can be used to assess candidate performance in a large number of examinations in India, which involves multiple choice questions. In India, most of the large-scale exams involving multiple choice questions are essential because they are used in government job recruitment and admission for higher studies. Since the latent traits of the IRTree model provide a viable alternative to the classical test score, it can be used singly or in conjunction with the classical test score in evaluating candidates in a large number of examinations in India.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., & Rcpp, L. (2015). Package 'lme4'. *Convergence*, *12*(1). https://cran.r-project.org/package=lme4

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443-459.

Böckenholt, U. (2013). Modeling multiple response processes in judgment and choice.

Böckenholt, U., & Meiser, T. (2017). Response style analysis with threshold and multi-process IRT models: A review and tutorial. *British journal of mathematical and statistical psychology*, *70*(1), 159-181.

Bolt, D. M., Wollack, J. A., & Suh, Y. (2012). Application of a multidimensional nested logit model to multiple-choice test items. *Psychometrika*, *77*(2), 339-357.

De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, *48*(1), 1-28. https://cran.r-project.org/package=irtrees

Debeer, D., Janssen, R., & De Boeck, P. (2017). Modeling Skipped and Not-Reached Items Using IRTrees. *Journal of Educational Measurement*, *54*(3), 333-363.

De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical education*, *44*(1), 109-117.

DeMars, C. (2010). *Item response theory*. Oxford University Press.

Ding, L., & Beichner, R. (2009). Approaches to data analysis of multiple-choice questions. *Physical Review Special Topics-Physics Education Research*, *5*(2), 020103.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and psychological measurement*, *58*(3), 357-381.

Glas, C. A., & Pimentel, J. L. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement*, *68*(6), 907-922.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.

Holman, R., & Glas, C. A. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology*, *58*(1), 1-17.

Jeon, M., & De Boeck, P. (2016). A generalized item response tree model for psychological assessments. *Behavior research methods*, *48*(3), 1070-1085.

Jeon, M., De Boeck, P., & van der Linden, W. (2017). Modeling answer change behavior: An application of a generalized item response tree model. *Journal of Educational and Behavioral Statistics*, *42*(4), 467-490.

Jeon, M., & Rijmen, F. (2016). A modular approach for item response theory modeling with the R package flirt. *Behavior research methods*, *48*(2), 742-755.

Köhler, C., Pohl, S., & Carstensen, C. H. (2015). Taking the missing propensity into account when estimating competence scores: Evaluation of item response theory models for nonignorable omissions. *Educational and psychological measurement*, *75*(5), 850-874.

Loken, E., & Rulison, K. L. (2010). Estimation of a four‐parameter item response theory model. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 509-525.

Lord, F. M. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika*, *39*(2), 247-264.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, *51*(2), 177-195.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series*, *1992*(1), i-30.

Nering, M. L., & Ostini, R. (Eds.). (2011). *Handbook of polytomous item response theory models*. Taylor & Francis.

Okumura, T. (2014). Empirical differences in omission tendency and reading ability in PISA: An application of tree-based item response models. *Educational and Psychological Measurement*, *74*(4), 611-626.

Posada, D., & Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic biology*, *53*(5), 793-808.

Rasch, G. (1960). Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In *Handbook of modern item response theory* (pp. 271-286). Springer, New York, NY.

Robitzsch, A., Kiefer, T., & Wu, M. (2019). TAM: Test analysis modules. R package version 3.1-45. https://CRAN.R-project.org/package=TAM

Stegmann, G., Jacobucci, R., Harring, J. R., & Grimm, K. J. (2018). Nonlinear Mixed-Effects Modeling Programs in R. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(1), 160-165.

Swaminathan, H., & Gifford, J. A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics*, *7*(3), 175-191.

Appendix A

R code for PCM and GPCM formulation

(Dataset used is in “final”)

Loading required package-

```
library(TAM)
```

Modeling-

PCM:

```
fit1 = tam.mml(final )
```

GPCM:

```
fit2 = tam.mml.2pl(final, irtmodel = "GPCM", control = list(conv=1E-9 ,
                acceleration = "none"))
```

Model summary and fit-statistics –

```
summary(fit1)

summary(fit2)
```

Ability Estimation:

```
ability1 = IRT.factor.scores(fit1,type = c("EAP"))

ability2 = IRT.factor.scores(fit2,type = c("EAP"))
```

R code for formulation of four IRTree models

(Dataset used is in “final”)

Loading required packages-

```
library(irtrees)

library(lme4)
```

For mapping:

```
linmap  =  cbind(c(1, 0, 1), c(0, NA, 1))
```

For transforming the data into long format:

```
linrespT = dendrify(final , linmap)
```

Model formulation:

```
Model1 = glmer(value ~ 0 + item:node + (0 + node | person), + family = binomial(link = "logit"), data = linrespT, nAGQ=0)

Model2 = glmer(value ~ 0 + item:node + (1 | person), + family = binomial(link = "logit"), data = linrespT, nAGQ=0)

Model3 = glmer(value ~ 0 + item + node + (1 | person), + family = binomial(link = "logit"), data = linrespT, nAGQ=0)

Model4 = glmer(value ~ 1 + (0 + node | item) + (0 + node | person), + family = binomial(link = "logit"), data = linrespT, nAGQ=0)
```

Model summary and fit-statistics –

```
summary(Model1)

summary(Model2)

summary(Model3)

summary(Model4)
```

Ability Estimation:

```
ability = ranef (Model1)
```

Item parameter Estimation:

```
item = fixef (Model1)
```

R code for formulation of three IRTree structures

For Model I mapping

```
linmap  =  cbind(c(1, 0, 1), c(0, NA, 1))
```

For Model II mapping

```
linmap  =  cbind(c(0, 1, 1), c(NA, 0, 1))
```

For Model III mapping

```
linmap  =  cbind(c(1, 1, 0), c(1, 0, NA))
```

(Rest of the code is unchanged)

Appendix B

Item difficulty values for Aerospace Engineering (AE) dataset

Table 9: Item difficulty values from IRTree model for AE dataset. Here -b$_1$ is interpreted as the intensity of omission induction while b$_2$ is interpreted as easiness (negative of difficulty).

Item number	b_1	StdError$_1$	b_2	StdError$_2$
itemi01	1.393616	0.043944	-2.28886	0.065259
itemi02	1.476186	0.04454	-0.07885	0.041599
itemi03	0.157619	0.040028	-0.90372	0.055555
itemi04	4.55082	0.126918	2.746057	0.063
itemi05	2.969899	0.066146	0.778653	0.039433
itemi06	2.673592	0.059869	1.46559	0.044209
itemi07	2.487359	0.05651	-1.09459	0.043369
itemi08	-0.28266	0.040549	-0.71919	0.058545
itemi09	1.218169	0.042837	-0.46108	0.043921
itemi10	1.650448	0.045961	-0.18381	0.041345
itemi11	1.316631	0.043433	0.8169	0.044198
itemi12	1.141376	0.042417	-0.99489	0.047827
itemi13	1.78165	0.047189	0.537997	0.041021
itemi14	2.050298	0.050162	0.317824	0.040151
itemi15	0.94377	0.041504	-0.0323	0.044128
itemi16	1.800298	0.047375	-0.80562	0.043401
itemi17	1.203534	0.042754	0.542583	0.04344
itemi18	3.201554	0.071947	1.008297	0.040509
itemi19	2.808302	0.062573	-0.62238	0.040086
itemi20	0.976443	0.041639	-0.62258	0.046153
itemi21	0.742225	0.040811	-1.27798	0.052933
itemi22	1.353307	0.043671	0.296052	0.042102
itemi23	2.078398	0.050511	-0.00948	0.039712
itemi24	1.563333	0.045222	-0.77189	0.044024
itemi25	1.29546	0.043299	1.227874	0.046988
itemi26	1.26698	0.043124	-0.91002	0.046581
itemi27	0.833989	0.041098	-0.78177	0.048123
itemi28	0.526187	0.040317	-0.88656	0.051409
itemi29	0.266643	0.040043	-0.75859	0.052898
itemi30	0.550169	0.040359	-2.05649	0.061159
itemi31	-0.55557	0.041332	-1.31032	0.072004
itemi32	0.532492	0.040328	-0.69809	0.048736

itemi33	0.793151	0.040964	-1.92693	0.063505
itemi34	0.891219	0.041301	0.58995	0.044815
itemi35	0.865844	0.041209	0.283246	0.044567
itemi36	-1.02591	0.043517	-0.16313	0.066108
itemi37	1.928086	0.04873	-0.51988	0.041517
itemi38	2.47333	0.056274	-0.68845	0.040901
itemi39	1.619309	0.04569	-0.96973	0.04479
itemi40	2.49018	0.056557	-0.365	0.039681
itemi41	2.07622	0.050484	-0.15167	0.040151
itemi42	2.065372	0.050349	-2.32339	0.059433
itemi43	2.498678	0.056702	-1.5567	0.047089
itemi44	1.591995	0.045459	-0.7361	0.04355
itemi45	1.926084	0.048707	-1.56061	0.049308
itemi46	1.507059	0.044775	-1.59596	0.051825
itemi47	1.460093	0.04442	-1.52859	0.050902
itemi48	1.110031	0.042256	-2.34653	0.066051
itemi49	0.903289	0.041346	-3.3616	0.098334
itemi50	1.364099	0.043743	-0.74394	0.044421
itemi51	1.559985	0.045195	-3.49437	0.0963
itemi52	1.962458	0.049119	-3.02666	0.078208
itemi53	0.581826	0.04042	-2.15347	0.067417
itemi54	0.774814	0.040907	-1.76318	0.057471
itemi55	1.046799	0.04195	-2.3107	0.064264
itemi56	1.087417	0.042144	-3.0936	0.084471
itemi57	0.559021	0.040376	-2.69969	0.085115
itemi58	0.541327	0.040343	-2.02323	0.064547
itemi59	0.291445	0.040055	-4.93557	0.216781
itemi60	0.970979	0.041616	-3.31781	0.092797
itemi61	1.393616	0.043944	-3.83909	0.114572
itemi62	0.457092	0.04021	-3.1707	0.104514
itemi63	2.084948	0.050594	-2.2785	0.05862
itemi64	0.493468	0.040263	-2.55567	0.07721
itemi65	1.653935	0.045992	-0.84367	0.043809