Federal University of Minas Gerais

School of Engineering

Department of Production Engineering

Graduate Program in Production Engineering

A Network Approach to Deal with the Problem of Examinee Choice under Item Response Theory

Thesis submitted by

Carolina Silva Pena

Under supervision of

Prof. Marcelo Azevedo Costa

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy in Production Engineering

Belo Horizonte

December 2016

P397n

Pena, Carolina Silva.
A network approach to deal whit the problem of examinee choice under item response theory [manuscrito] / Carolina Silva Pena. – 2016.
ix, 75 f., enc.: il.

Orientador: Marcelo Azevedo Costa.

Tese (doutorado) - Universidade Federal de Minas Gerais, Escola de Engenharia.

Bibliografia: f. 69-75.

1. Engenharia de produção - Teses. 2. PERT (Análise de redes) - Teses. 3. Teoria da resposta do item - Teses. 4. Teoria bayesiana de decisão estatística I. Costa, Marcelo Azevedo. II. Universidade Federal de Minas Gerais. Escola de Engenharia. III. Título.

CDU: 658.5(043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

UFMG

FOLHA DE APROVAÇÃO

A Network Approach to Deal with the Problem of Examinee Choice under Item Response Theory

CAROLINA SILVA PENA

Tese submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em ENGENHARIA DE PRODUÇÃO, como requisito para obtenção do grau de Doutor em ENGENHARIA DE PRODUÇÃO, área de concentração PESQUISA OPERACIONAL E ENGENHARIA DE MANUFATURA, linha de pesquisa Modelagem Estocástica e Simulação.

Aprovada em 06 de dezembro de 2016, pela banca constituída pelos membros:

Marcelo Azevedo Costa

Prof(a). Marcelo Azevedo Costa - Orientador
UFMG

Erica C Rodrigues

Prof(a). Erica Castilho Rodrigues
ICEB/UFOP

Prof(a). Allbens Atman Picardi Faria
CEFET-MG

Prof(a). Marcos Antonio da Cunha Santos
DEST/UFMG

Prof(a). Martin Gomez Ravetti
UFMG

Belo Horizonte, 6 de dezembro de 2016.

Acknowledgments

I would like to express my gratitude to Professor Marcelo Azevedo Costa, who was my Academic Advisor during the masters' degree and now in the doctorate program. He revised my project several times, gave me valuable suggestions and, above all, helped me to develop as a researcher. I realized my evolution during the writing of the second paper. The first paper was very difficult to me, mainly because of the language barrier and the several rules to be addressed during the submission process. The second paper flowed much more naturally. I can say that Professor Marcelo taught me how to write a manuscript for a top international journal.

I would like to thank my beloved husband, Rivert, for encouraging me many times during this academic journey and also for valuable contributions, in the second paper, with his experience in Bayesian modeling. My special thanks to my beloved family, my mother Dora, my father Deusdedit and my brothers Guilherme and Will. After all, they have always stood by me in this world, which provided me confidence to achieve this academic degree. I dedicate this work to my friends and family.

I would like to thank the support of Graduate Studies Office of the Federal University of Minas Gerais, especially Professor Walmir Matos Caminhas and Professor Ricardo Hiroshi Caldeira Takahashi.

Resumo

Em uma situação típica de avaliação utilizando questões de prova, os candidatos não podem escolher quais itens preferem responder. A principal razão é um problema técnico em se obter estimativas confiáveis para as habilidades dos candidatos e as dificuldades dos itens. Este trabalho propõe uma nova representação dos dados utilizando análise de redes. As questões de prova, e os itens selecionados, para cada candidato, são codificados como camadas, vértices e arestas em uma rede multicamadas. Dessa forma, um novo modelo de Teoria de Resposta ao Item (TRI), que incorpora a informação obtida a partir da rede multicamadas utilizando modelagem Bayesiana, é proposto. Diversos estudos de simulação, nos quais os candidatos podem escolher um subconjunto de itens, foram realizados. Os resultados mostram uma melhora substancial na recuperação dos parâmetros utilizando o modelo proposto em comparação ao modelo convencional de TRI. Este modelo é a primeira proposta que permite obter estimativas satisfatórias em cenários críticos reportados na literatura.

Palavras-chaves: Seleção de itens, Análise de Redes, Medidas de Centralidade, Teoria de Resposta ao Item, Modelagem Bayesiana.

Abstract

In a typical questionnaire testing situation, the examinees are not allowed to choose which items they would rather answer. The main reason is a technical issue in obtaining satisfactory statistical estimates of examinees' abilities and items' difficulties. This paper introduces a new Item Response Theory (IRT) model that incorporates information from a novel representation of the questionnaire data, using network analysis. The questionnaire data set is coded as layers, the items are coded as nodes and the selected items are connected by edges. The new proposed Item Response Theory (IRT) model incorporates network information using Bayesian estimation. Several simulation studies in which examinees are allowed to select a subset of items were performed. Results show substantial improvements in the parameters' recovery over the standard model. To the best of our knowledge, this is the first proposal to obtaining satisfactory IRT statistical estimates in some critical scenarios reported in literature.

Keywords: Examinee choice, Selection of items, Network analysis, Centrality Measures, Item Response Theory, Bayesian Modeling.

Summary

List of Figures

List of Tables

Chapter 1

Introduction

One important challenge faced by large scale tests is to provide comparability between examinees performance, particularly in assessment situations in which it is not possible to apply the same test for them all. Furthermore, the comparability of the examinees performance over the years is crucial for evaluating the educational system. This problem has been overcome by estimating the examinees abilities using a set of statistical models that compose the Item Response Theory (IRT). This property, knowing as invariance, is obtained in IRT models by introducing separate parameters for examinees and items. The IRT models provides optimal estimates when all items within a test are mandatory to the examinees. On the other hand, if the examinees are allowed to choose a subset of items, instead of answer them all, the model estimates can became seriously biased. As a consequence, examinees typically are not allowed to choose which items they would rather answer.

The problem of allowing examinee choice has been raised by many works [Wainer, Wang and Thissen, 1994; Fitzpatrick and Yen, 1995; Wang, Wainer and Thissen, 1995; Bradlow and Thomas, 1998; Linn, Betebenner and Betebenner, 1998; Cankar, 2010; Wang et al., 2012], but remains without a proposal that can satisfactorily deal with it. This thesis is comprised by two manuscripts that address such context. The first manuscript, i.e., the Chapter 2, explore a novel representation of the data set using network analysis. The examinees and their select items are coded as layers, nodes and edges. Two matrices are obtained from this representation. Moreover, several network centrality measures are calculated using both matrices in order to find a statistical measure that strongly correlates with items difficulties.

The second manuscript, i.e., the Chapter 3, introduces a new IRT model that incorporates the information obtained from network analysis by using Bayesian estimation. Results show substantial improvements in the parameters' recovery over the standard Rasch model. To the best of our knowledge, this is the first proposal to obtaining satisfactory IRT statistical estimates in some critical scenarios reported in literature.

Chapter 2

A Network Approach to Evaluate the Problem of Examinee Choice under Item Response Theory

Abstract

Item Response Theory (IRT) models are widely applied to estimate examinee ability and item difficulty using questionnaire data. However, if examinees are allowed to choose fewer items in the questionnaire then statistical estimates of parameters can be biased. This is because the choice may depend on the examinee personal ability and on the item difficulty. In this work, we explore a novel representation of examinees and their selected items using a multilayer network: the questionnaire data is coded as layers, nodes and edges. Using the multilayer network, we propose a statistical method to create a monolayer network from which network centrality measures are calculated. Simulated and real case data sets show that the estimated centrality measures are strongly and statistically correlated to item difficulty. In addition, we propose a statistical inference test to determine if the examinees are randomly selecting items. Findings are currently being investigated to minimize statistical bias in IRT estimates.

Keywords: Network analysis, Network aggregation, Centrality measures, Item Response Theory, Examinee Choice, Missing data.

Introduction

This study analyzes the problem of evaluating individuals (or examinees) in a specific field of interest using a questionnaire or test. The test has V items and for each item each individual must provide a response. After evaluating the completed test, a value is given to each item. Let $Y_{i\alpha}$ be the random variable of interest that represents the value for each item i ($i = 1, \ldots, V$) and examinee α ($\alpha = 1, \ldots, M$). If the responses of the examinees are correct then $Y_{i\alpha} = 1$, otherwise $Y_{i\alpha} = 0$. Classical Test Theory uses the number of correct responses, $\sum_i Y_{i\alpha}$, to estimate the examinee ability. However, this approach does not take into account the fact that different items have different levels of difficulty [Hambleton, Swaminathan and Rogers, 1991]. For example, suppose two examinees achieve the same number of correct responses but the subset of items answered by the first examinee are, in general, more difficult. In this example, the first examinee has a superior ability which is not identified only by using the number of correct items. In practice, the examinee ability, hereafter identified as θ_α, and the item difficulty, hereafter identified as b_i, are not known in advance. They must be estimated using the set of responses $Y_{i\alpha}$. Item Response Theory (IRT) is the methodology which is widely used to estimate examinee ability and item difficulty [Hambleton, Swaminathan and Rogers, 1991].

IRT has optimal statistical properties when responding to all V items is mandatory for all examinees. On the other hand, if the examinees are free to choose v items ($v < V$) from the total number of items (V), then the model estimates can be statistically biased since the choice may depend on the examinee ability and on the item difficulty. In general, the examinees will choose the v items for which the difficulties are most closely matched to their personal abilities [Wang, Wainer and Thissen, 1995; Bradlow and Thomas, 1998; Wang et al., 2012]. Nonetheless, if a particular subset of items is frequently chosen, then this subset of items can be evaluated as if it were mandatory. Thus, statistical estimates of the abilities and difficulties can potentially be improved. Furthermore, if the subset of items, hereafter named non-random subset, can be identified then the frequencies of correct responses for each item can be used to differentiate more difficult items from less difficult items. Therefore, it is of interest to identify summary statistics which can be used to select potential mandatory subsets, and which are potentially correlated to the difficulties of the items.

In this work, we explore a novel representation of the examinees and their selected items using a multilayer network. Each examinee is represented as a single-layer network in which the total number of items are the vertices, or nodes, and the selected items are fully connected by edges. That is, each edge connects pairs of items chosen by the examinee. Therefore, each layer of the multilayer network represents one examinee and the selected items. This novel representation may show relationships among the examinees and items which are not identified using only frequency of selection.

Many systems can be seen as networks or multilayer networks. For example, social networks [Verbrugge, 1979; Barrett, Henzi and Lusseau, 2012], gene co-expression networks [Li et al., 2011], transportation networks [Cardillo et al., 2013], climate networks [Donges et al., 2011], among others [Kivelä et al., 2014]. Furthermore, a multilayer representation may include interaction between layers, generally represented as connecting edges between layers. Multilayer network analysis implies that relevant information might not be identified if the single layers were analyzed separately [Menichetti et al., 2014; Battiston, Nicosia and Latora, 2014]. On the other hand, summary statistics easily can be estimated from a monolayer network. Therefore, it is of interest to represent multilayer networks as a compact layer from which relevant information regarding the problem is obtained [De Domenico et al., 2015].

Our proposed methodology summarizes the questionnaire multilayer as follows: first we propose a single weighted network in which, for each edge, the weight is related to the frequency of the selected pairs of items; next, we propose a statistical procedure to eliminate edges which are found to be randomly selected. Thus, the final network is composed of edges which are statistically evaluated as non-random. In addition, we propose a Monte Carlo statistical inference procedure to test the null hypothesis that the examinees are choosing items randomly. This is a useful test for investigating whether the layers in a multilayer network are correlated [Bianconi, 2013]. Finally, the following centrality measures: betweenness, closeness, degree, eigenvector and strength are calculated using the monolayer network in order to find statistical measures which are correlated to item difficulty. Simulated and real case studies show that the proposed analysis provides relevant information about item difficulty, which potentially can be used to improve IRT estimates.

Figure 2.1 illustrates our proposed network representation considering data from 8 examinees (A-H). Each examinee selected 5 items from a questionnaire with a total of 20 items (nodes). All items selected by one examinee are a clique of connected nodes [Luce and Perry, 1949]. Figures 2.1A to 2.1H show the examinee network layer. Figure 2.1I shows the summary network in which one edge represents a pair of items which was selected by at least one examinee [Battiston, Nicosia and Latora, 2014]. It can be shown that the larger the number of examinees the more saturated with edges the summary network eventually becomes.

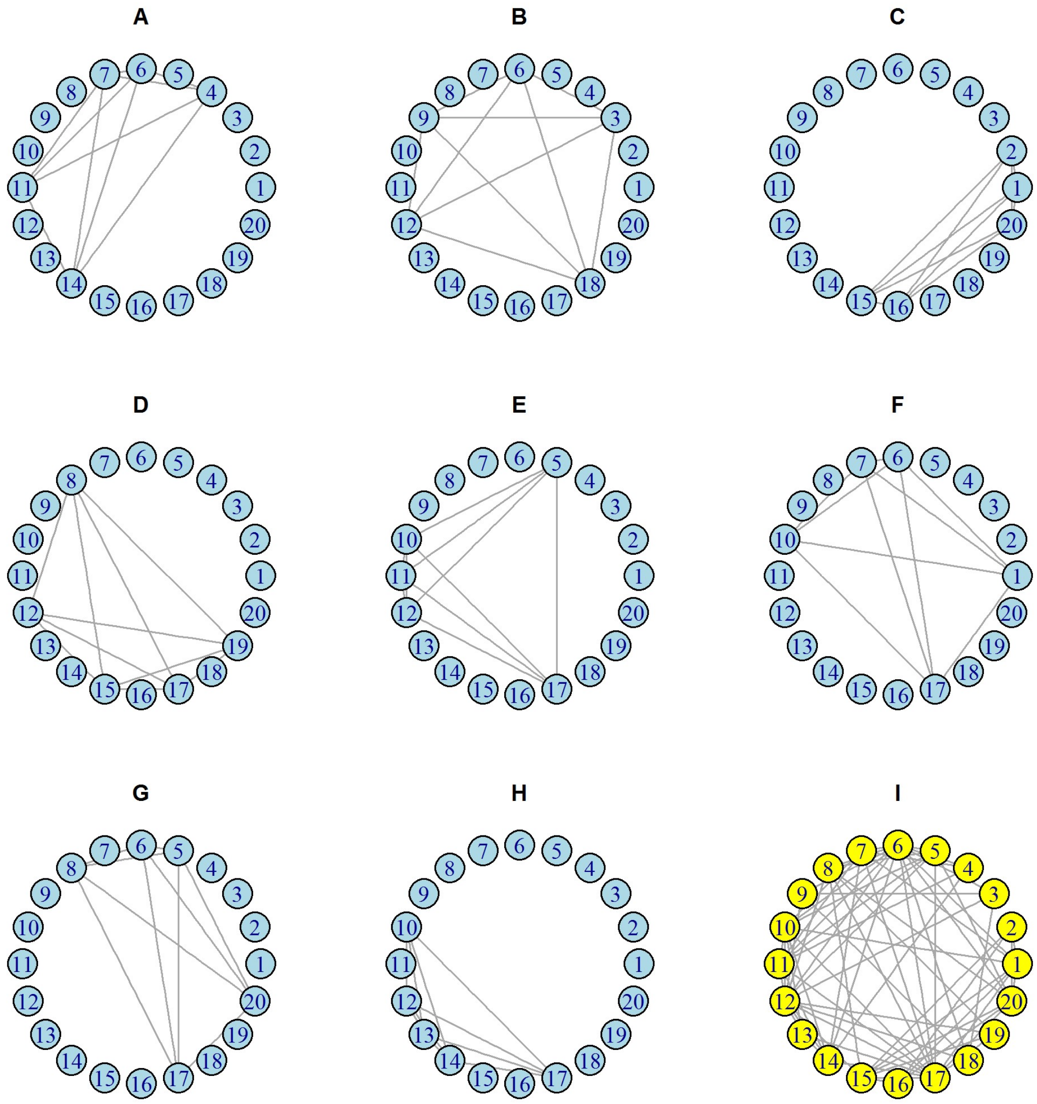

Figure 2.1: Network representation of selected items from a questionnaire with 20 items. (A-H) Represents 8 examinee networks, each with 5 selected items. In these networks, each vertex represents an item and all vertices selected by an examinee are fully connected. (I) Aggregated single-layer network in which each edge connects two items that were selected by at least one examinee.

Material and Methods

Following Barrat, Barthélemy and Vespignani [2008], a network is any system that admits an abstract mathematical representation using graphs. A graph $\mathrm{G} = \left(\vec{\mathrm{V}}, \vec{\mathrm{E}}\right)$ is a collection of V vertices (or nodes) which identifies the elements of a particular system and a set of E edges, related to the vector $\vec{E}$. Each edge connects pairs of vertices $\{\mathrm{v_i}, \mathrm{v_j}\}$, $\mathrm{v_i}, \mathrm{v_j} \in \vec{V}$ indicating the presence of a relationship or interaction between them. The maximum number of edges in network G is given by:

$$\eta_{\varepsilon} = \frac{V(V-1)}{2}, \tag{2.1}$$

and its density is defined by [Lewis, 2009, p. 53]:

$$\mathcal{D} = \frac{E}{\eta_{\varepsilon}} . \tag{2.2}$$

Network Aggregation

Consider a complex system represented using V vertices and M layers. Let $\vec{G} = (G_1, G_2, \dots, G_M)$ be a set of networks (or graphs) and G_α is the network at layer α, $\alpha = 1, \dots, M$. $\vec{G}$ is also known as a multilayer network [Battiston, Nicosia and Latora, 2014]. Each network, $G_\alpha = \left(\vec{V}, \vec{E}_\alpha\right)$, is represented using the vector of vertices $\vec{V}$ and the vector of edges $\vec{E}_\alpha$. Hereafter, it is assumed that the set of vertices $\vec{V}$ is the same for every layer α, while the set of edges $\vec{E}_\alpha$ depends on each layer.

Each network can be represented using an adjacency squared matrix $\mathbf{A}^{[\alpha]}$ of dimension V where $a_{ij}^{[\alpha]}$ = 1 if there is an edge between vertices i and j in layer α, and $a_{ij}^{[\alpha]}$ = 0, otherwise. Suppose each layer G_α has V_α selected vertices ($0 < V_\alpha \leq V$), which are fully connected by edges. Thus, the sum of the total number of edges present in the multilayer network $\vec{G}$ is:

$$E^{[o]} = \sum_{\alpha=1}^{M} \frac{V_\alpha(V_\alpha - 1)}{2} . \tag{2.3}$$

The simplest way to aggregate multiple layers is using aggregation matrices. For instance, let $\mathbf{A}$ be the aggregation matrix $\mathbf{A} = [a_{ij}]$, where:

$$a_{ij} = \begin{cases} 1, & if \; \exists \; \alpha \colon a_{ij}^{[\alpha]} = 1; \\ 0, & otherwise. \end{cases} \tag{2.4}$$

That is, pairs of vertices in matrix $\mathbf{A}$ are connected if there is, at least, one layer α in which vertices $\mathrm{v_i}$ and $\mathrm{v_j}$ are connected ($a_{ij}^{[\alpha]} = 1$) [Battiston, Nicosia and Latora, 2014]. Therefore, matrix $\mathbf{A}$ is a summary matrix of the multilayer $\vec{\mathrm{G}}$ which ignores multi-ties between pairs of nodes among layers.

The overlapping matrix $\mathbf{O} = [o_{ij}]$ is an alternative aggregation matrix in which the multi-ties between pairs of nodes are not ignored [Battiston, Nicosia and Latora, 2014; Bianconi, 2013; Barigozzi, Fagiolo and Garlaschelli, 2010]:

$$o_{ij} = \sum_{\alpha} a_{ij}^{[\alpha]}, \tag{2.5}$$

therefore $0 \leq o_{ij} \leq M \;\; \forall \, i, j$. The overlapping matrix $\mathbf{O}$ preserves the number of layers in which the connections are present as compared to matrix $\mathbf{A}$. Nevertheless, matrices $\mathbf{A}$ and $\mathbf{O}$ do not provide additional information about the existence of connections between layers, i.e., the inter-layer edges [Battiston, Nicosia and Latora, 2014].

Moreover, two layers of a multilayer network can be either dependent or independent [Bianconi, 2013]. A multilayer is independent if specific connections between nodes in one layer do not give any additional information about the chance of connecting nodes in different layers. On the contrary, dependent multilayer has recurrent connections among layers or inter-layer edges. This work proposes statistical tests to identify recurrent connections, and to indentify if a multilayer is dependent or independent. To do so, a new aggregation matrix, named matrix $\mathbf{U} = [u_{ij}]$, is proposed as follows:

$$u_{ij} = \begin{cases} 1, & if\ o_{ij} > \Psi;\ i \neq j; \\ 0, & otherwise. \end{cases} \tag{2.6}$$

where o_{ij} is given by equation 2.5 and Ψ is a critical upper bound value which is estimated under the null hypothesis (H_0) that the multilayer is independent. Let o_{ij} be the random variable of interest. It represents the number of times one edge between vertex $\mathrm{v_i}$ (hereafter known as vertex i) and vertex $\mathrm{v_j}$ (hereafter known as vertex j) is observed in the multilayer $\vec{G}$. Under the null hypothesis that the multilayer is independent, it can be shown that the expected number of times that one edge connecting vertices i and j happens in a multilayer is:

$$e_{ij} = \frac{E^{[o]}}{\eta_\varepsilon}. \tag{2.7}$$

The aggregation matrix $\mathbf{U}$ assumes that $u_{ij} = 1$ if there is evidence that the null hypothesis is rejected, or $o_{ij} > e_{ij}$. It is straightforward to show that:

$$o_{ij} > e_{ij} \rightarrow \quad o_{ij} > \frac{E^{[o]}}{\eta_\varepsilon} \rightarrow \frac{o_{ij}}{E^{[o]}} > \frac{1}{\eta_\varepsilon}. \tag{2.8}$$

Alternatively, let π_{ij} be a random variable defined as $\pi_{ij} = \frac{o_{ij}}{E^{[o]}}$. π_{ij} is the proportion of connections between vertices i and j among the sum of the total number of edges of the multilayer. Therefore, the null and alternative hypothesis are written as:

$$\begin{cases} H_0: \pi_{ij} = \pi_\varepsilon, \\ H_1: \pi_{ij} > \pi_\varepsilon. \end{cases} \tag{2.9}$$

Where $\pi_\varepsilon = \frac{1}{\eta_\varepsilon}$. The null and alternative hypothesis, described in equation 2.9, can be evaluated using a standard statistical proportion test [Casella and Berger, 2002, p. 493 - 494]. Let γ be the level of significance of the statistical test. It can be shown that the null hypothesis H_0 is rejected if:

$$o_{ij} > \frac{E^{[o]}}{\eta_\varepsilon} + Z_\gamma \sqrt{\frac{E^{[o]}}{\eta_\varepsilon}(1 - \frac{1}{\eta_\varepsilon})} \tag{2.10}$$

where Z_γ is z-score statistic. For example, if $\gamma = 0.05$ (5%) then $Z_\gamma = 1.645$. Therefore, given γ, $\Psi = \frac{E^{[o]}}{\eta_\varepsilon} + Z_\gamma \sqrt{\frac{E^{[o]}}{\eta_\varepsilon}(1 - \frac{1}{\eta_\varepsilon})}$. Alternativelly, Ψ is the upper bound of the observed number of edges between vertices i and j if $E^{[o]}$ edges were randomly distributed among the layers in the multilayer $\vec{G}$. It is worth mentioning that, under the null hypothesis, the statistical distribution of the number of incident edges in each vertex is the same for all vertices in the network. Consequently, the threshold is the same for all vertices. Thus, the proposed aggregation matrix $\mathbf{U}$ is a summary matrix in which vertices i and j are connected if the observed number of edges between vertices i and j in the multilayer $\vec{G}$ is statistically significant.

Figure 2.2A shows the overlapping network $G_{\mathbf{O}}$, which is represented by matrix $\mathbf{O}$ and Figure 2.2B the statistically significant network $G_{\mathbf{U}}$, which is represented by matrix $\mathbf{U}$. The simulated data was previously used in Figure 2.1. The value of Ψ was estimated as 1.49, using $Z_\gamma = 1.645$.

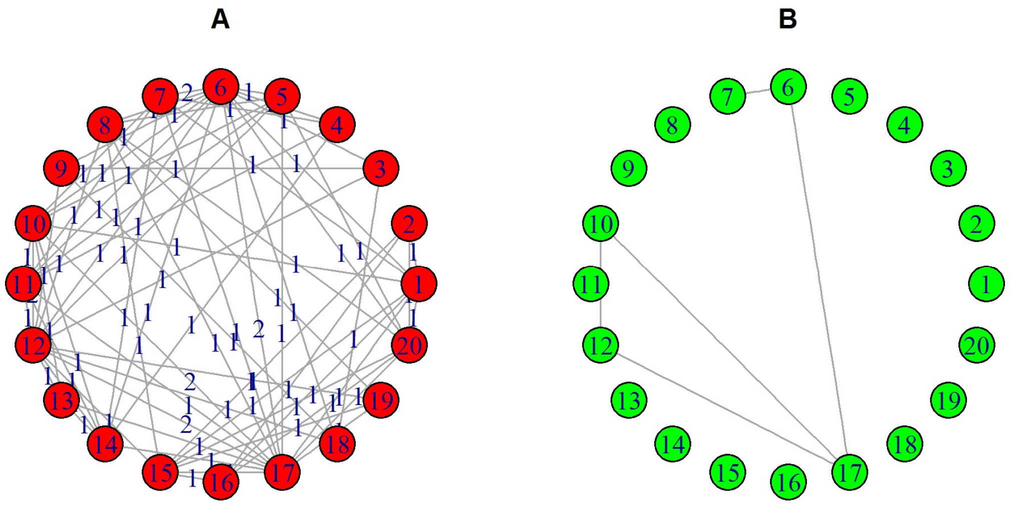

Figure 2.2: Network aggregation. (A) The overlapping network represented by matrix **O**. (B) The statistically significant network represented by matrix **U**.

Centrality Measures

As previously shown, the elements of the adjacency matrix $\mathbf{U}$ represent edges which are statistically proven to be non-random. This connecting structure among vertices can be further used to identify those which are the most connected. In our proposed questionnaire network representation, statistical measures of connections among vertices, known as centrality measures, are used to identify important vertices.

Some of the most frequently used centrality measures found in the literature [Batool and Niazi, 2014] include: betweenness centrality, closeness centrality, degree centrality, eigenvector centrality and strength centrality [Barrat et al., 2004]. The first three measures were proposed by Freeman [1978] and the eigenvector centrality was proposed by Bonacich [1972].

The degree of vertex i (Cd_i), given an adjacency matrix $\mathbf{U}$, is the number of first order neighbors:

$$Cd_i = \sum_{j \neq \mathrm{i}} u_{ij}, \tag{2.11}$$

where $u_{ij} \in \{0,1\}$.

The closeness centrality measure of vertex i, shown in equation 2.12, is the inverse of the sum of the distances from vertex i to all vertices in the network. The distance between two vertices i and j, l_{ij}, is the number of edges in the shortest path to reach vertex j starting from vertex i.

$$Cc_i = \frac{1}{\sum_{j \neq i} l_{ij}}. \tag{2.12}$$

If there is no path between two vertices (i. e., $l_{ij} = \infty$) the total number of vertices is generally used as the distance [Csardi and Nepusz, 2006].

The betweenness centrality of vertex i is the number of times the vertex i is found in the shortest paths between two other vertices. Let T_{hj} be the total number of shortest paths between vertices h and j, then $T_{hj}(i)$ is the number of those paths that include vertex i. The betweenness centrality of vertex i is:

$$Cb_i = \sum_{h \neq j \neq i} \frac{T_{hj}(i)}{T_{hj}}. \tag{2.13}$$

The eigenvector centrality of vertex i is the *i-th* element of the first eigenvector of the adjacency matrix $\mathbf{U}$**,**

$$\lambda \mathbf{x} = \mathbf{U}\mathbf{x}, \tag{2.14}$$

where λ is the eigenvalue and $\mathbf{x}$ is the first eigenvector of matrix $\mathbf{U}$. The eigenvector centrality is a relative score assigned to each vertex. In general, vertices with high eigenvector centralities are those which are connected to many other vertices which are, in turn, connected to many others (and so on). Further information about eigenvector centrality score is found in Bonacich [1972].

The strength of a vertex i, given the overlapping matrix **O**, is the sum of the weights of all the edges incident on a vertex i:

$$Cs_i = \sum_j o_{ij}. \tag{2.15}$$

Multilayer Network Statistical Independence Test

The edges represented in matrix $\mathbf{U}$ rejected the null hypothesis shown in equation 2.9. Given statistical significance level γ, a certain number of the edges will reject the null hypothesis even if the null hypothesis is true. This is known as the error type I [Casella and Berger, 2002, p. 382 - 383]. Therefore, given the number of edges in matrix $\mathbf{U}$, we propose a global statistical test using the number of edges of matrix $\mathbf{U}$ or, similarly, the density of matrix $\mathbf{U}$ defined in equation 2.2. A Monte Carlo inference is proposed [Kroese, Taimre and Botev, 2011, p. 281- 343].

Under the null hypothesis that the multilayer is independent and given the total number of edges in each layer, a multilayer network is simulated by randomly connecting pairs of vertices in each layer. Then, matrix $\mathbf{U}$ and its density $\mathcal{D}$ are estimated. This procedure is repeated S times, say $S = 10{,}000$, to generate an empirical distribution of $\boldsymbol{\mathcal{D}} = \mathcal{D}_1, \ldots, \mathcal{D}_{10{,}000}$. The final p-value is calculated as the proportion of simulated densities greater than the observed density ($\mathcal{D}_{obs}$):

$$P\ value = \sum_{s=1}^{10{,}000} I(\mathcal{D}_s \geq \mathcal{D}_{obs})/10{,}000 \tag{2.16}$$

where $I(\mathcal{D}_s \geq \mathcal{D}_{obs})$ = 1 if $\mathcal{D}_s \geq \mathcal{D}_{obs}$ and 0, otherwise. If the p-value $< \gamma$, then it can be concluded that the null hypothesis is false.

Fundamentals of Item Response Theory

Item Response Theory (IRT) is a psychometric theory used to evaluate data collected from a questionnaire or test. It is used to estimate the abilities of the examinees and also the difficulties of the questions (or items). The examinee ability estimated by IRT do not depend on the questionnaire, i.e., examinees estimated abilities can be compared even if different questionnaires were applied. This property, known as invariance, is one important property of IRT as compared to classical test theory [Hambleton and Swaminathan, 1985]. The invariance property is obtained by introducing statistical models with separate parameters for the examinee ability and the item difficulty [Hambleton, Swaminathan and Rogers, 1991, p. 2 - 8]. The Rasch model [Rasch, 1960] is a widely used IRT model. The item characteristic curve for the Rash model is given by [Hambleton, Swaminathan and Rogers, 1991, p. 13]:

$$P(Y_{i\alpha} = 1|\theta_\alpha, b_i) = \frac{e^{\theta_\alpha - b_i}}{1+e^{\theta_\alpha - b_i}} , \tag{2.17}$$

where $P(Y_{i\alpha} = 1|\theta_\alpha, b_i)$ is the probability model that a randomly chosen examinee with ability θ_α correctly answers item i. In the Rasch model, the parameter b_i represents the ability required for any examinee to have a 50% (0.50) chance of correctly answering item i. Estimates for θ_α and b_i in the IRT models are found using the following likelihood equation [Hambleton, Swaminathan and Rogers, 1991]:

$$f(\vec{y}|\vec{\theta}, \vec{b}) = \prod_{\alpha=1}^{M} \prod_{i=1}^{V} (P(Y_{i\alpha} = 1|\theta_\alpha, b_i))^{y_{i\alpha}} (1 - P(Y_{i\alpha} = 1|b_i, \theta_\alpha))^{1-y_{i\alpha}}, \tag{2.18}$$

where $\vec{\theta} = (\theta_1, \theta_2, \dots, \theta_M)$ is the vector that contains the M examinees abilities, $\vec{b} = (b_1, b_2, \dots, b_V)$ is the vector that contains the V difficulties of the items and $\vec{Y}$ is the vector of observed responses.

When allowing examinee choice, the likelihood equation includes the missing-data-indicator vector $\vec{R} = (R_{1,1}, R_{1,2}, \dots, R_{i,\alpha}, \dots, R_{V,\mathrm{M}})$ where $R_{i\alpha} = 1$ if the examinee α response for item i is observed; otherwise, $R_{i\alpha} = 0$ [Bradlow and Thomas, 1998]. Given $\vec{R}$, the set of responses $\vec{Y}$ can be written as $\vec{Y} = (\vec{\mathrm{Y}}_{obs}, \vec{\mathrm{Y}}_{mis})$ [Rubin, 1976], where $\vec{\mathrm{Y}}_{obs}$ denotes the observed values and $\vec{\mathrm{Y}}_{\mathrm{mis}}$ denotes the missing values. Thus, the new likelihood function, described in Bradlow and Thomas [1998], is given by:

$$f\left[\left(\vec{\mathrm{Y}}_{obs}\,,\vec{\mathrm{Y}}_{mis}\right),\vec{R}\middle|\vec{\theta},\vec{b}\right]=f\left[\vec{R}\middle|\left(\vec{\mathrm{Y}}_{obs}\,,\ \vec{\mathrm{Y}}_{mis}\right),\vec{\theta},\vec{b}\right]f\left[\left(\vec{\mathrm{Y}}_{obs}\,,\ \vec{\mathrm{Y}}_{mis}\right)\middle|\vec{\theta},\vec{b}\right]. \quad (2.19)$$

Bradlow and Thomas [1998] showed that if the examinees are allowed to choose items then valid inference for $\vec{\theta}$ and $\vec{b}$ can be obtained, using equation (2.18), only if the following assumptions are applied:

$$f\left[\vec{R}\middle|\left(\vec{\mathrm{Y}}_{obs}\,,\ \vec{\mathrm{Y}}_{mis}\right),\vec{\theta},\vec{b}\right]=f\left[\vec{R}\middle|\vec{\mathrm{Y}}_{obs},\vec{\theta},\vec{b}\right], \quad (2.20)$$

$$f\left[\vec{R}\middle|\vec{\mathrm{Y}}_{obs},\vec{\theta},\vec{b}\right]=f\left[\vec{R}\middle|\vec{\mathrm{Y}}_{obs}\right]. \quad (2.21)$$

The first assumption (2.20) is known as the missing at random (MAR) assumption and implies that examinees are not able to distinguish items that they would usually find difficult, given their abilities. The second assumption (2.21) implies that examinees of different abilities generally do not select broadly the easier or the more difficult items. Details can be found in Bradlow and Thomas [1998]. Hereafter, it is assumed that if the examinees are not randomly selecting items then the unobserved values are not missing at random.

To further illustrate the consequences of the two assumptions described above, suppose that item i is very difficult, therefore, only a few examinees with high ability choose to answer this item. In this example, the estimates of the difficulty parameter for item i will possibly be underestimated. This is because the abilities of the examinees are unknown and only a few data for the item i are available. Wang et al. [2012] proposed to include an additional random effect parameter to account the process of the selection of items. Nevertheless, the proposed model was not able to produce valid statistical inference using the simulation study described in Bradlow and Thomas [1998], in which the MAR assumption is false.

Particularly, if the IRT data is structured as aggregation networks then frequencies and correlations among items (or vertices) selected by the examinees can be represented using centrality measures. Furthermore, equation 2.6 creates the aggregation matrix $\mathbf{U}$ whose edges were individually evaluated under the null hypothesis that the assumptions described in Bradlow and Thomas [1998] are true. That is, matrix $\mathbf{U}$ represents the questionnaire post-processed network in which the edges represent pairs of items that rejected the null hypothesis of MAR. In addition, centrality measures can be evaluated as potential predictors of the item difficulty.

Simulated and real data sets are used to estimate the elements of matrix $\mathbf{U}$ and to evaluate statistical correlations among the centrality measures and item difficulty.

Simulation Study

The proposed network aggregation method and the centrality measures were evaluated using three simulated scenarios. In all scenarios, there is an additional random-effect parameter γ_α for each examinee that account for the choice effect, hereafter known as choice parameter [Wang et al., 2012]. All scenarios have a questionnaire with 50 items ($i = 1, \dots, 50$) and 1,000 examinees ($\alpha = 1, \dots, 1{,}000$). The examinees can choose 20 items. In all scenarios, the abilities of the examinees, θ_α, and the difficulties of the items, b_i, are generated using a normal density distribution with mean zero and variance one, $\theta_\alpha \sim Normal(\mu = 0;\ \sigma^2 = 1)$ and $b_i \sim Normal(\mu = 0;\ \sigma^2 = 1)$.

The first scenario, hereafter named scenario 1, is based on the simulation study presented by Bradlow and Thomas [1998]. In this scenario, the choice parameter γ_α was set equal to the personal ability ($\gamma_\alpha = \theta_\alpha$). As previously mentioned, the probability of an examinee correctly answered item i using the Rasch model is greater than 0.50 (50%) if $\theta_\alpha > b_i$. Therefore, for each examinee α, the items are divided into two groups: the first group comprises items which are easier as compared to the examinee ability, i.e., $b_i \leq \gamma_\alpha$. This is the group in which the examinee has a probability higher than 0.50 to achieve a correct answer. The second group comprises items which are more difficult as compared to the examinee ability, i.e., $b_i > \gamma_\alpha$. In this group, the examinee has a probability lower than 0.50 to achieve a correct answer. Next, a weight value (w_i) is assigned to each item. For items in group 2, the weight value is $w_i^{[2]} = 1$; whereas, for items in group 1, the weight value varies from 1 to 2: $w_i^{[1]} \in \{1, 1.1, 1.2, 1.3, \ldots, 1.9, 2.0\}$. If $w_i^{[1]} = 1$ then the MAR assumption is true; whereas, for $w_i^{[1]} > 1$, the MAR assumption is false. For example, if $w_i^{[1]} = 1.1$, then it can be said that the items in group 1 have a 10% higher chance of being selected as compared to the items in group 2. For each weight value in group 1, $w_i^{[1]}$, 10,000 Monte Carlo simulations are used. Furthermore, for each Monte Carlo simulation, each examinee chooses 20 items. Selected items are generated using a multinomial probability distribution. That is, the probability of examinee α selecting item i is:

$$p_{i\alpha} = \frac{w_{i\alpha}}{\sum_i w_{i\alpha}}. \tag{2.22}$$

In the second scenario (scenario 2), similarly to scenario 1, the items are divided into two groups: bellow and above the choice parameter γ_α. However in the scenario 2 the choice parameter is not fixed, it is updated after each choice made by the examinee, i. e., the choice parameter was defined as a vector $\gamma_{\alpha t}$, where t = 1, ..., 20. The selection of the first item is similar to scenario 1, i.e., $\gamma_{\alpha 1} = \theta_\alpha$. Suppose the first selected item has a difficulty value of $b_{(1)}$. Then, $\gamma_{\alpha 2} = b_{(1)}$ and the remaining items will be divided into two new groups using the following rule: one item belongs to group 1 if $b_i \leq \gamma_{\alpha 2}$ and belongs to group 2 if $b_i > \gamma_{\alpha 2}$. Given the weigth value for group 2, the next item (second item) is selected using the multinomial distribution and equation 2.22. After choosing the second item, the items will be divided into two new groups again based on the average value of items difficulty level, previously selected, or $\gamma_{\alpha t} = \sum_{k=1}^{t-1} b_{(k)}/(t-1)$. Therefore, the groups of easy and difficult items are changed using the rule: one item belongs to group 1 if $b_i \leq \gamma_{\alpha t}$ and belongs to group 2 if $b_i > \gamma_{\alpha t}$. Thus, the probabilities of selecting new items are changed using the past selected items.

The proposed second scenario is assumed to be the more consistent with a questionnaire applied over a longer period of time. Suppose, for instance, undergraduate programs of 4 to 5 years in which an examinee (or student) may pass or fail a course (or item) in each academic semester. As the examinee passes or fails some of the items, the choice parameter is changed. Consequently, the next items (courses) are chosen based on cumulative experience. In the proposed scenario, the choice parameter is estimated using the difficulty values of the past selected items. This scenario assumes that the tendency of examinees to try harder items may increase or decrease over time.

The third scenario, named scenario 3, is based on a job selection trainee program. In this scenario, the job candidates or the examinees are free to choose 20 items from a total number of 50 items. Suppose that each examinee must choose one item after a training period. The examinees must compete among themselves in order to finish the questionnaire with the largest number of difficult items with correct responses. In this situation, the difficulty of the selected item must be closer to the ability of the examinee. Examinees which select the much more difficult items are more likely to provide incorrect responses and fail the trainee program. Similarly, examinees which select the easiest items may also fail the trainee program. Given the previous description, the following simulation, named scenario 3, is proposed: initially, items with similar difficulties are grouped using the *k-means* clustering method [Hartigan and Wong, 1979]. A total of 10 groups is created. Each group has items with similar difficulties and different groups may have different numbers of items. On average, each group has 5 items. In the first step we set $\gamma_{\alpha 1} = \theta_\alpha$ and each examinee is more likely to choose items from the closest group of items using the minimum distance between the choice parameter

$\gamma_{\alpha 1}$ and the average group difficulty. In addition to the closest group, each examinee is also more likely to choose items from adjacent groups: the closest group with either a slightly higher or slightly lower average difficulty. These three groups are hereafter named the target group. Different from scenarios 1 and 2, in which the examinees are more likely to choose the easiest items, this scenario creates groups of items from which each examinee is more likely to choose items. Each simulation uses different weights for the items in the target group, $w_i^{[1]} > 1$; whereas for items in the remaining groups, $w_i^{[2]} = 1$. As a consequence, the probability of the examinees choosing items from the target group is greater than the probability of them choosing items from the remaining groups. Similar to scenario 2, after each selected item, each choice parameter ($\gamma_{\alpha t}$) is updated using the 60% percentile of the difficulties of the previously selected items. Table 2.1 shows the main features of each simulated scenario; also, the differences between them.

Table 2.1: Summary of the main features of the three simulated scenarios

Scenario	**Choice parameter**	**Group of items to which greater weight is assigned**
1	Fixed	Items with difficulty below the choice parameter
2	Changed according to previous choice	Items with difficulty below the choice parameter
3	Changed according to previous choice	Items with difficulty close to the choice parameter

It is worth mentioning that in each simulated scenario 10,000 Monte Carlo simulations were evaluated using $w_i^{[1]} = 1$. These are the simulations in which the MAR assumption is valid. That is, the selection of items is random, regardless of the abilities of the examinees or the difficulties of the items. As a consequence, the simulated multilayer networks are independent and both assumptions described in Bradlow and Thomas [1998] are true. Furthermore, different intensities of dependence, i.e., the MAR assumption is false, were evaluated: $w_i^{[1]} \in \{1, 1.1, 1.2, 1.3, \dots, 1.9, 2.0\}$ for scenarios 1 and 2 and for scenario 3,

$w_i^{[1]} \in \{1, 1.5, 2.5, 5.0, 10, 20, 30, 50\}$. This is because, for scenario 3, convergence of the network density was achieved for larger values of $w_i^{[1]}$, as compared to scenarios 1 and 2. The simulation study was performed using the R software [R Core Team, 2015].

Case Studies

The School of Engineer Data Set

The data set comprises 23 elective (i.e., not mandatory) subjects attended by 217 students of Control and Automation Engineering in years 2004, 2005 and 2006, at the Federal University of Minas Gerais (UFMG), Brazil. We obtained permission from the Graduate Studies Office of the Federal University of Minas Gerais to access this data set in compliance with Brazilian Law no. 12527 which states that all information produced or held in the custody of the government, except for personal or legally classified information, is public and therefore accessible to all citizens. No personal identification of the students was required, i.e., the data was de-identified.

The selected 217 students were enrolled in at least two of the 23 subjects. Most of the students attended from 6 to 12 courses. Therefore, the associated multilayer network has 217 layers, each layer represents the student network of selected courses. The total number of edges in the multilayer network ($E^{[o]}$) is 10,253. The total number of vertices, in each layer, is 23.

In addition to the selected courses, the difficulty level of each course (b_i) was estimated using a survey. The coordinator of the undergraduate program assigned a score from 1 to 10 for each subject, where 1 means very easy and 10 means very difficult. In order to include the evaluations by the students, the questionnaire was put online. To request the participation of the students, a professor from the undergraduate program (cited in the acknowledgments) sent an email to all students. Completion of the questionnaire was entirely voluntary. The data was anonymous, i.e., there was no field in the questionnaire for personal identification. Twenty four former students assigned score values to the courses. The linear correlation of median score values of the former students and the coordinator score values is 0.62. The final estimate of the difficulty level for each course is the average value of the coordinator score and the median score of the former students.

The coordinator also classified each course into one of the following groups: low relevant course, medium relevant course and high relevant course. Each group is based on the opinion of the coordinator about the course contribution to the minimum preparation of a control and automation engineer.

The Brazilian Lottery Data Set

The *Mega Sena* is a very popular lottery game in Brazil. It happens two times a week, on average. The game comprises 6 numbers which are randomly selected from a group of 60 numbers: $1, 2, \ldots, 60$. Selected numbers from previous games are available online (http://loterias.caixa.gov.br/wps/portal/loterias/landing/megasena). All numbers drawn in the Brazilian Lottery are obtained by clicking on the link "*Resultado da Mega sena por ordem crescente*" located at the bottom of the page. The data are publicly available for free download, therefore no permission was required.

We evaluated the numbers from the first game, in March 11, 1996 to game number 1,741, which happened in September 12, 2015. Therefore, the multilayer network has 1,741 layers. Each layer has a total of 60 vertices and 6 selected vertices, forming a clique. The total number of edges in the multilayer network ($E^{[o]}$) is 26,115. This data set was selected because it represents a real case data in which the MAR assumption is supposed to be true.

Results

Figure 2.3 shows boxplots of the density statistic ($\mathcal{D}$) of matrix $\mathbf{U}$ for the simulated scenarios (1, 2 and 3), for the different values of w_i. Note that if $w_i = 1$ then the MAR assumption is true. Figure 2.3 also shows a horizontal line which represents the 95th percentile of the simulated densities for $w_i = 1$. Therefore, simulations in which the density is above the horizontal line statistically reject the MAR assumption. For scenario 1, shown in Figure 2.3A, when $w_i = 1.1$ the simulated density distribution was slightly different from the MAR simulated density. If $w_i > 1.1$ then the simulated density distributions become much different from the MAR simulations. In general, the larger the weigth w_i the greater the differences between simulated densities and simulated densities under the MAR assumption. Therefore, simulated results for scenario 1 show that the greater the density statistic of matrix **U** the more likely the examinees are choosing easier items given their personal abilities.

Figure 2.3B shows simulated results for scenario 2. It is worth mentioning that the results for $w_i = 1$ are similar to scenario 1. The greater the value of w_i the greater the differences between simulated densities and the MAR simulations, as similar to scenario 1. Furthermore, scenario 2 has a larger increasing rate of the distances as compared to scenario 1. For example, the median density for scenario 1 is 0.082 if $w_i = 1.2$, whereas for scenario 2 the median density is 0.132 using the same w_i value. This is because selected items are affected by previous selected items in scenario 2. At each step, when the examinee chooses an easier item, the number of items that comprises group 1 became smaller. As a consequence, the chance of choosing these fewer (and easier) items in the future is high. Thus, the network density rate is larger for scenario 2 as compared to the network density rate for scenario 1.

Figure 2.3C shows simulated results for scenario 3. Larger values for w_i were used to show convergence of the density statistic, which achieved convergence for smaller values of w_i

in scenarios 1 and 2. Similar to scenarios 1 and 2, the larger the values of w_i the larger the density statistic. Furthermore, the density statistic achieved larger values before reaching convergence, as compared to scenarios 1 and 2. This is because the edges in matrix **U**, in scenario 3, are not concentrated toward the easiest items.

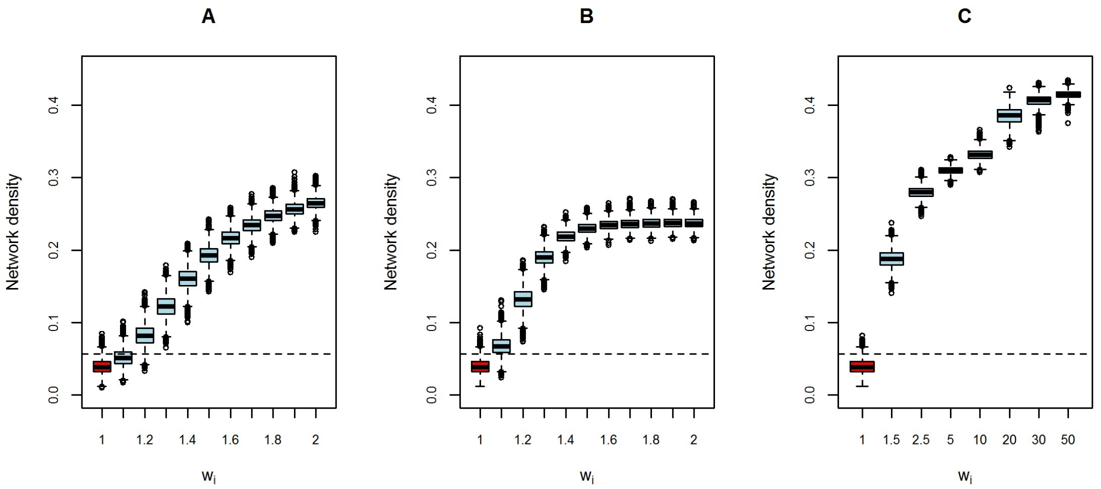

Figure 2.3: Boxplots of the simulated densities for different values of weight w_i. (A) Scenario 1. (B) Scenario 2. (C) Scenario 3. The lower outer contour of the rectangle indicates the first quartile (Q1), the upper outer contour of the rectangle indicates the third quartile (Q3) and the horizontal line inside the rectangle indicates the median (Q2). Vertical lines extended from the box indicate variability outside the first and third quartiles. The upper vertical line indicates the maximum observed value within the range [Q3; Q3+1.5×(Q3-Q1)]. The lower vertical line indicates the minimum observed value within the range [Q1; Q1-1.5×(Q3-Q1)]. Observations beyond the vertical lines are represented as points and indicate outliers.

In addition to the density statistic, the centrality measures were also evaluated for simulated scenarios 1, 2 and 3. For different values of w_i, the Pearson linear correlation coefficient [McDonald, 2014, p. 190 – 208] between simulated difficulties (b_i) and estimated centrality measures for each item was calculated. Figures 2.4 - 2.9 show mean values of the linear correlation and the 95% highest probability density (HPD)[Chen and Shao, 1999] interval for each w_i value. Figure 2.4, Figure 2.6 and Figure 2.8 show centrality measures calculated using matrix **O**. Figure 2.5, Figure 2.7 and Figure 2.9 show centrality measures calculated using matrix **U**. It is worth noticing that the HPD intervals are centered at zero if $w_i = 1$. This is because under MAR assumption the correlation between the difficulties of the items and centrality measures are zero, on average.

Results show that the linear correlation between item difficulty and degree centrality, and between item difficulty and the betweenness centrality, using matrix **O**, was always zero. This is because, in all simulations, matrix **O** became fully connected and, consequently, both degree and betweenness centrality measures assumed the same value for all vertices. It is worth mentioning that the correlation between these two centrality measures and the difficulty of the items were set to zero because there is no correlation if one of the variables is actually a constant. In scenarios 1 and 2, the closeness centrality measure using matrix **O** is positively correlated to item difficulty. This is because the weights of the edges are interpreted by the algorithm as cost [Dijkstra, 1959]. Therefore, the easier items, with smaller values of b_i, also present smaller values of closeness, and vice-versa. If matrix **U** is used, then the correlation is negative. This is because matrix **U** is binary and sparser and, as a consequence, the easiest items have larger values of closeness, and vice-versa. Furthermore, using matrix **O,** HPD intervals for strength and eigenvector centrality measures are very similar for scenarios 1 and 2. For scenario 3, the HPD intervals for the eigenvector centrality measure are superior, with higher mean and non-overlapping intervals, as compared to the HPD interval of the strength centrality measure. It is worth noticing that since the examinees choose a fixed number of items (20 items), the strength centrality measure is proportional to the frequency of selected items. As a conclusion, for scenarios 1, 2 and 3 the eigenvector centrality measure is consistently one of the centrality measures most correlated to item difficulty. Therefore, there is evidence that the eigenvector centrality measure is a superior and robust statistic for evaluating item difficulty as compared to the evaluated centrality measures.

Unlike matrix **O**, for a larger number of examinees, matrix **U** does not become saturated. As a consequence, none of the evaluated centrality measures become saturated. Furthermore, the strength and degree centrality measures are similar if matrix **U** is used. This is because the strength centrality measure, using a non-weighted matrix, is equal to the degree centrality measure. It is worth noticing that using matrix **U** the closeness and betweenness centrality measures, which are related to network paths, did not correlate equally well with item difficulty as compared to degree and eigenvector centrality measures. This is because matrix **U** is a simplified matrix in which the edges are related to the strength of the items rather than to network internal connectivity. Therefore, no physical meaning was assigned to the distance between two items in matrix **U**.

In general, the eigenvector centrality measure using matrix **O** achieved slightly superior results as compared to matrix **U** for scenarios 1 and 2. This is because the easiest items are the most likely choices among the examinees in these scenarios. These items are easily detected using frequency information, which is best represented in matrix **O**. It is worth mentioning that matrix **U** is a sparser matrix carrying partial information of matrix **O**. Nonetheless, results using matrix **U** are very close to the results using matrix **O**. Thus, there is evidence that the information loss using matrix **U** is minimal.

For scenario 3, results using matrix **U** are slighlty superior than using matrix **O**. This is because the examinees are more likely to choose different groups of items according to their personal abilities, as opposed to a commom group, which are the easiest items used in previous scenarios. In general, examinees in scenario 3 are more likely to choose items from the target group. The target group has some items which are sliglty more difficult or less difficult as compared to the examinee ability. If the examinees choose the easier items in the target group then confounding elements are created in matrix **O**. The filtering procedure applied to matrix **O** minimizes the presence of confounding elements in matrix **U**. Therefore, centrality measures using matrix **U** are more correlated to item difficulty as compared to matrix **O** in scenario 3.

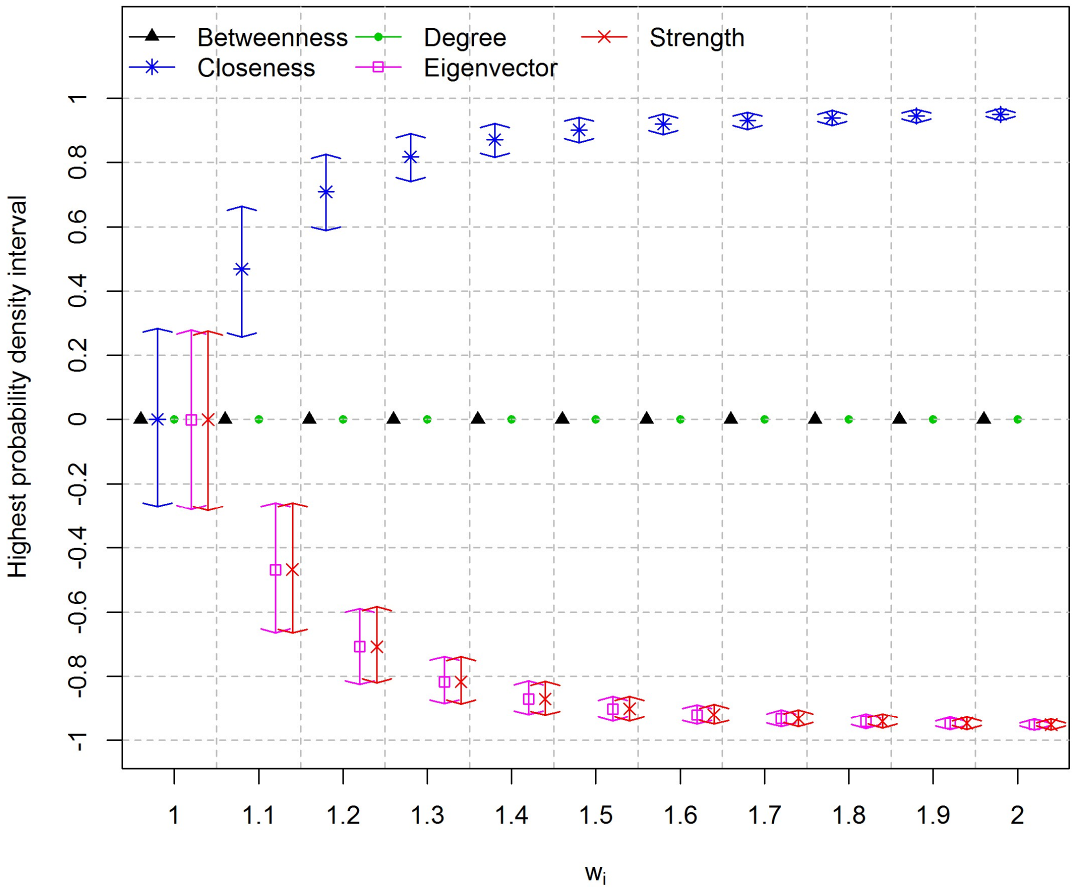

Figure 2.4: HPD intervals for Pearson linear correlation coefficient between simulated difficulties and centrality measures using matrix O for different weight values (w_i) – Scenario 1. HPD is the narrowest interval containing 95% of the Monte Carlo estimates. The symbols inside the intervals indicates the mean value of all the Monte Carlo estimates.

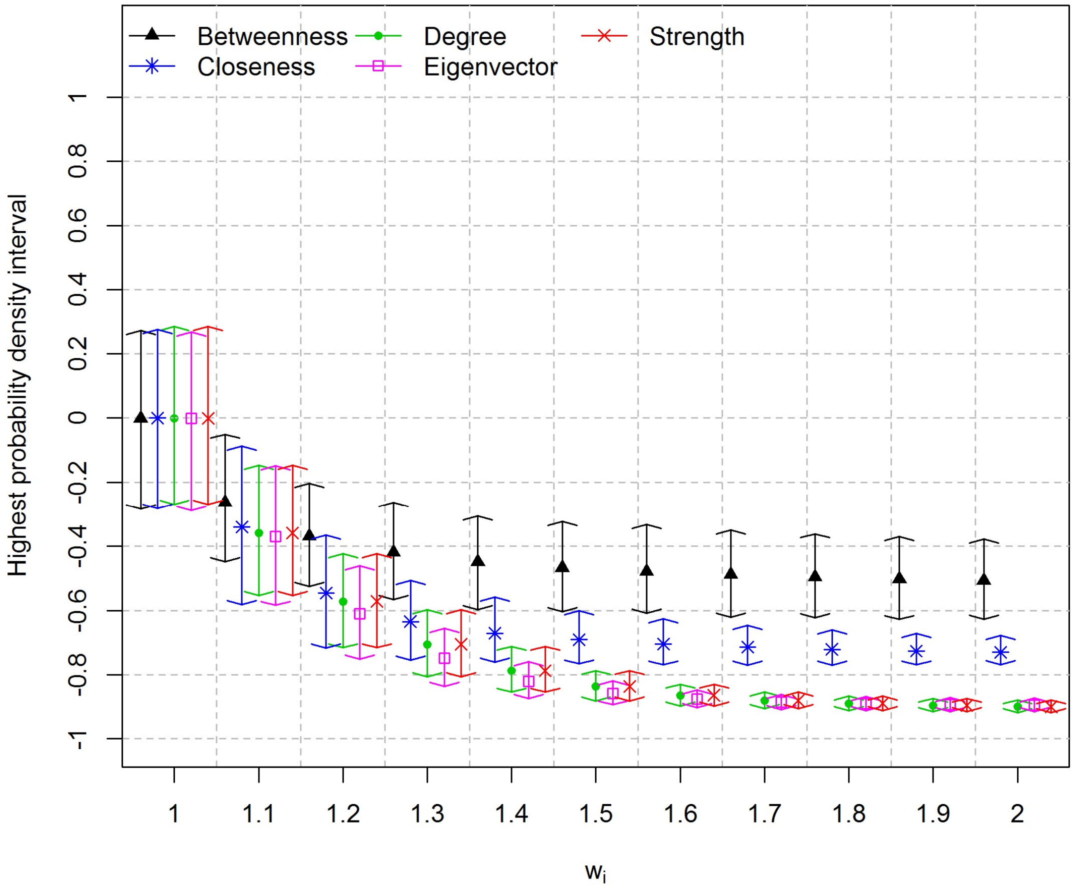

Figure 2.5: HPD intervals for Pearson linear correlation coefficient between simulated difficulties and centrality measures using matrix U for different weight values (w_i) – Scenario 1.

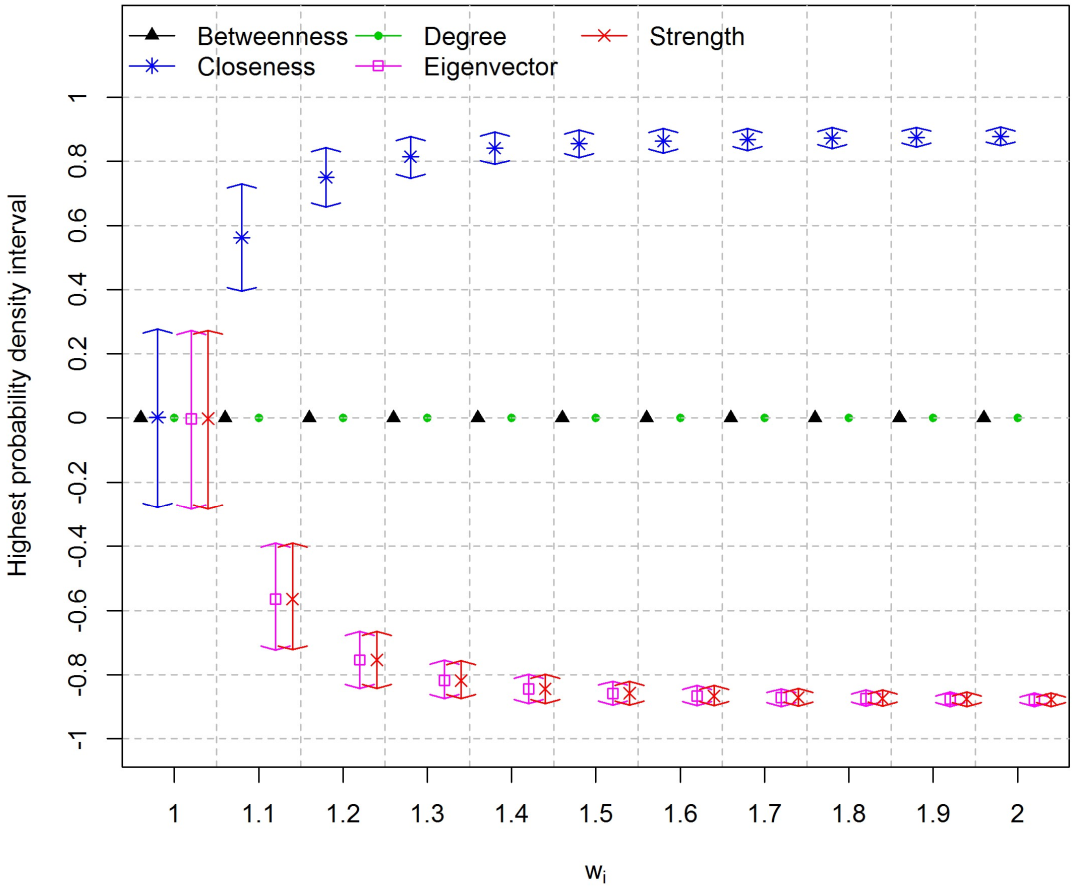

Figure 2.6: HPD intervals for Pearson linear correlation coefficient between simulated difficulties and centrality measures using matrix O for different weight values ($\mathrm{w_i}$) – Scenario 2.

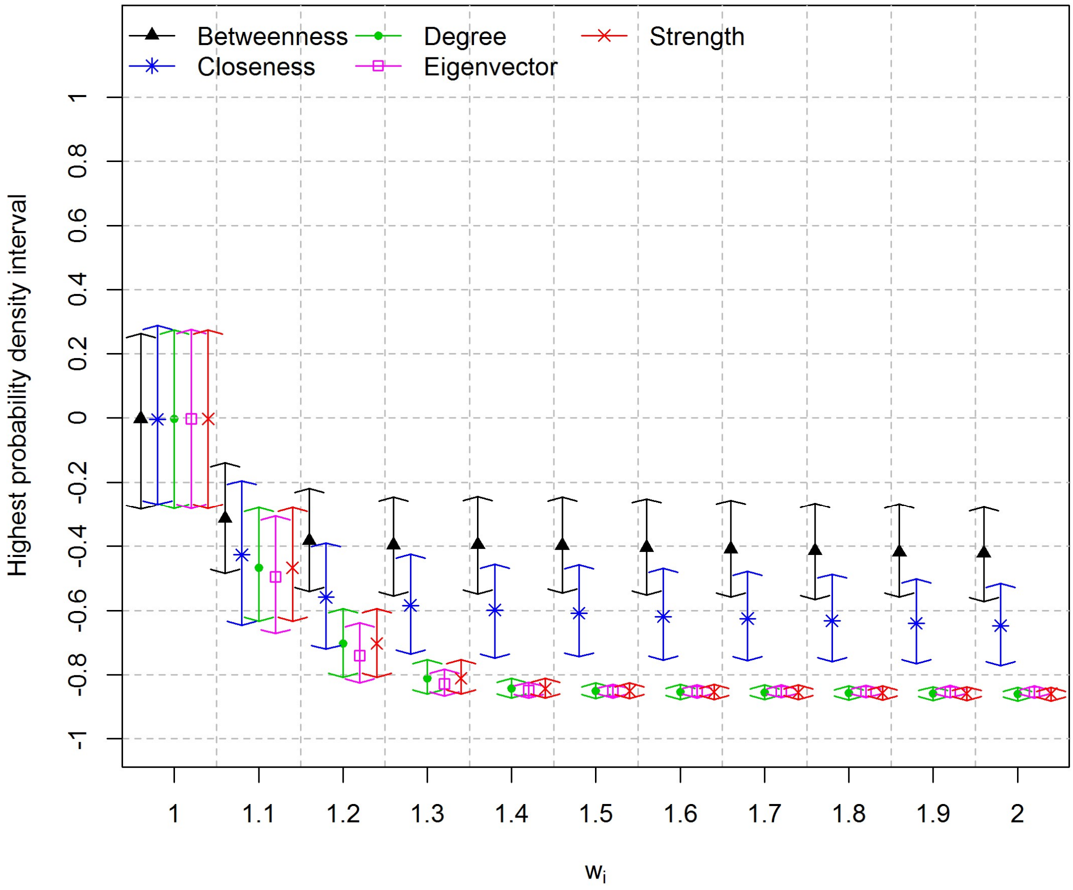

Figure 2.7: HPD intervals for Pearson linear correlation coefficient between simulated difficulties and centrality measures using matrix U for different weight values (w_i) – Scenario 2.

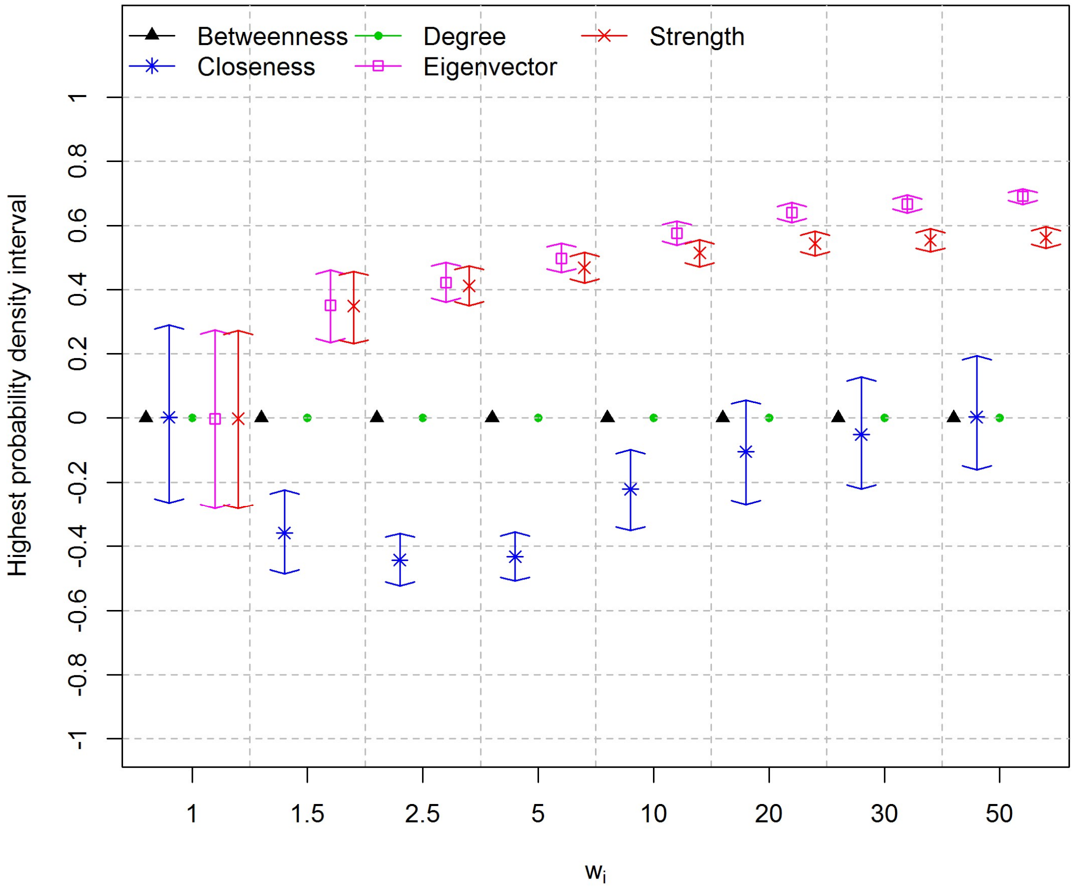

Figure 2.8: HPD intervals for Pearson linear correlation coefficient between simulated difficulties and centrality measures using matrix O for different weight values (w_i) – Scenario 3.

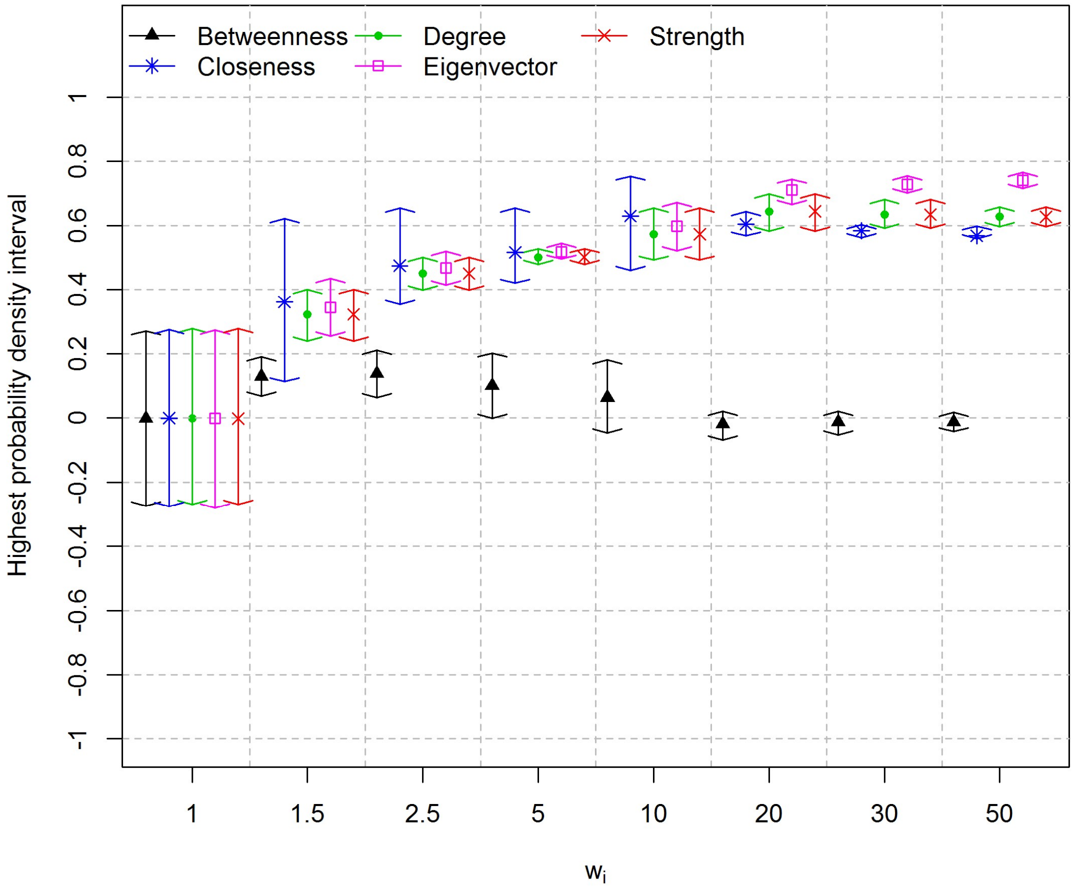

Figure 2.9: HPD intervals for Pearson linear correlation coefficient between simulated difficulties and centrality measures using matrix U for different weight values (w_i) – Scenario 3.

Finally, Figure 2.10 shows the bias and the estimated eigenvector, using matrix **O**, both when the MAR assumption is valid (Figure 2.10A) and for different violations of the MAR assumption (Figure 2.10B to Figure 2.10D), using scenario 1. Difficulties were estimated using marginal maximum likelihood [Bock and Aitkin, 1981]. The bias of the estimates is the difference between true simulated item difficulty and estimated item difficulty, $b_i - \hat{b}_i$. In addition, information regarding whether the degree of matrix **U** is equal to zero or greater than zero is included. If the degree of matrix **U** is equal to zero, then the items were selected by fewer examinees and were considered as non-recurrent items in the multilayer network analysis. If the degree of matrix **U** is greater than zero, then the items were statistically evaluated as recurrent items in the multilayer network. It is worth noticing that the eigenvector is a measure within the range 0–1. If the MAR assumption is not violated, then the eigenvector is concentrated towards 1, as shown in Figure 2.10A. The more violated the MAR assumption is, i.e. the greater the w_i, the larger the range of the eigenvector. Therefore, the range of the estimated eigenvector is a measure which indicates the degree of MAR violation. Furthermore, for lower values of w_i, the variability of the bias is very similar to the MAR assumption. Therefore, it can be concluded that, for mild to moderate violations of the MAR assumption, the bias of the estimates is similar to the bias of the MAR estimate, as shown in Figure 2.10B and Figure 2.10C. The stronger the violation of the MAR assumption, the closer to zero is the lower range of the eigenvector and the larger the bias variability, as shown in Figure 2.10D. Figure 2.10 also shows that items with larger eigenvector values are associated with items having degrees greater than zero. Therefore, these items were frequently selected by the examinees and, consequently, the bias is lower as compared to items with smaller eigenvectors and degrees equal to zero. These new results show strong evidence that the eigenvector centrality measure is correlated with the bias of the estimates of item difficulty and is a standardized statistic, since it lies within the range 0–1. Very similar results were found for scenarios 2 and 3: the more violated the MAR assumption, the larger the range of the eigenvectors; and, the lower the eigenvector, the greater the variability of the bias.

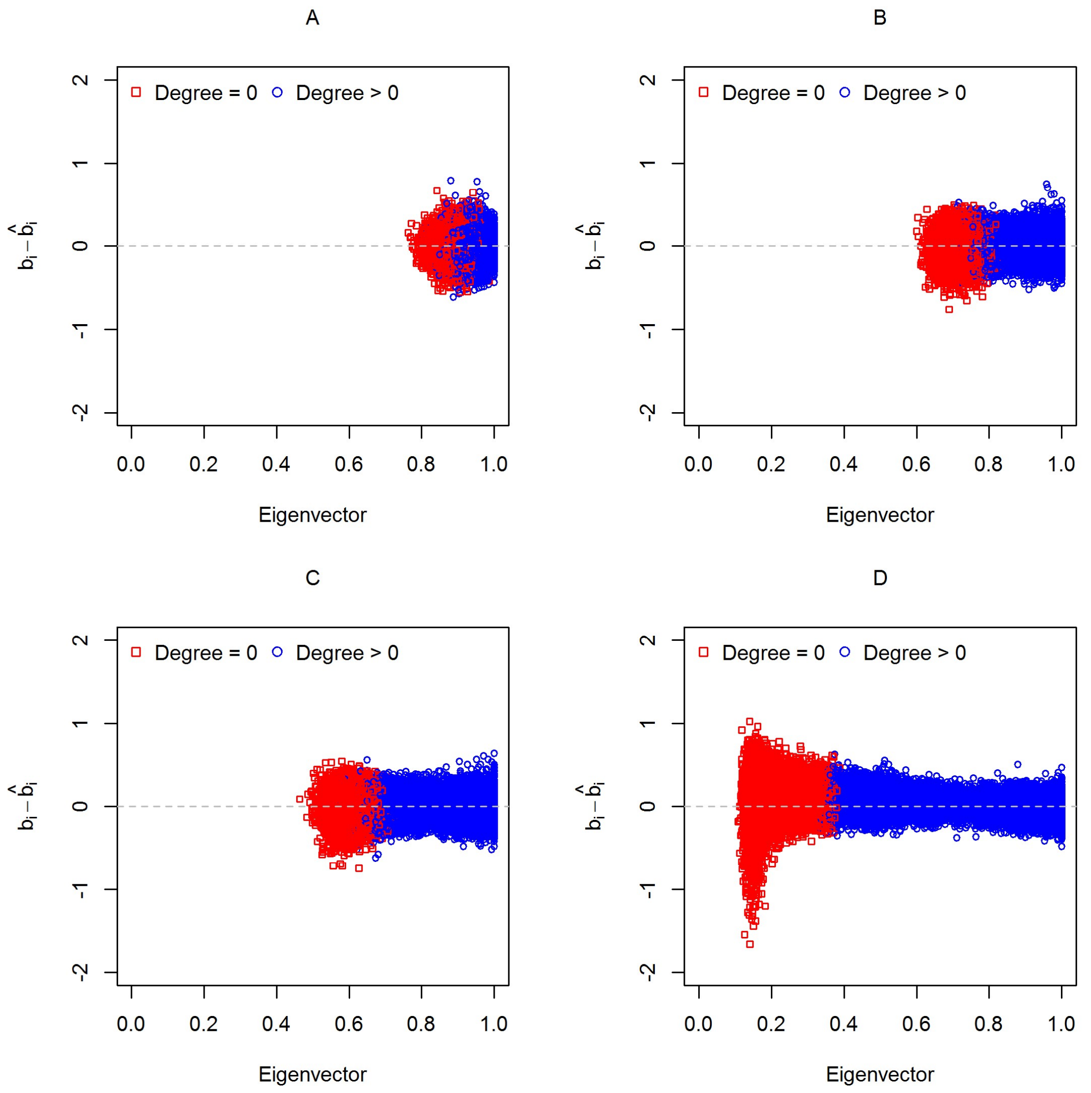

Figure 2.10: Bias of estimated item difficulty versus eigenvector centrality measure, using matrix O, for simulated scenario 1. The legend shows two groups of items: those with degree equal to zero and those with degree greater than zero, using matrix **U**. (A) The MAR assumption is true. (B) Weight value is 1.5. (C) Weight value is 2. (D) Weight value is 20.

Case Studies

Results: School of Engineering Data Set

Figure 2.11A shows the overlapping network of the school of engineering data set. The critical upper bound value was estimated as $\Psi = 50.98$, using a statistical significance level of

5%. Thus, matrix $\mathbf{U}$ has two connected subjects (or vertices) if, at least, 51 students selected the same pair of subjects. Figure 2.11B shows the final network related to matrix $\mathbf{U}$. The network has 17 connected vertices, or subjects. The size of the vertex represents the difficulty level of each subject, i.e., the larger the vertex, the larger the estimated difficulty score.

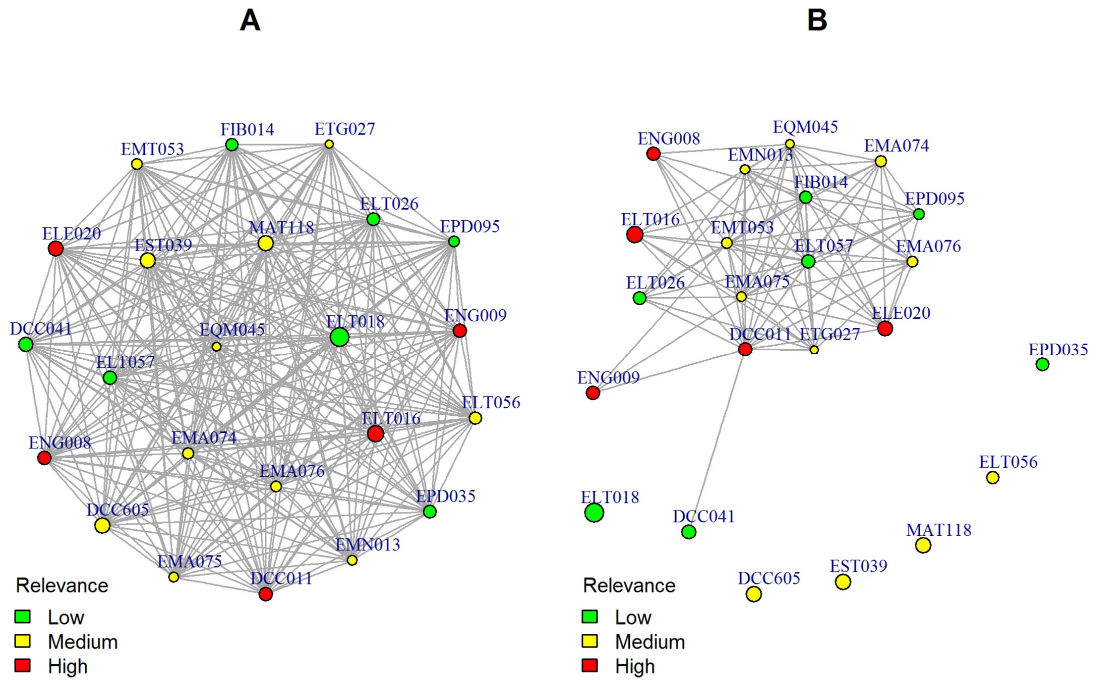

Figure 2.11: School of engineering network. (A) Single-layer network using the aggregation matrix $\mathbf{A}$. (B) Single-layer network using the aggregation matrix $\mathbf{U}$. Vertex size is proportional to the difficulty level of each subject. Vertex color indicates the relevance of the course to the minimum preparation of a control and automation engineer, according to the opinion of the coordinator.

The p-value, estimated using equation 2.16, is less than 0.0001 ($< 10^{-4}$). Therefore, there is statistical evidence that the null hypothesis of an independent multilayer network is false, i.e., there is statistical evidence of a group of subjects commonly selected by the students, shown in Figure 2.11B.

In addition, the Pearson linear correlation coefficient between eigenvector centrality from matrix **U** and estimated difficulties; and between strength centrality from matrix **O** and estimated difficulties were estimated, as shown in Table 2.2. The correlation between eigenvector measures and estimated difficulties is -0.62 (p-value = 0.0015), and between strength measures and estimated difficulties is -0.57 (p-value =0.0048). For low and high relevant subjects the correlations are statistically non-significant. For high relevant subjects the difficulty does not seem to be a determining factor in the student decision, i. e., the correlation coefficient is close to zero. Furthermore, there are fewer subjects in low and high relevant subjects' groups. On the contrary, for medium relevant subjects the correlation can be considered statistically significant. This indicates that the students are selecting subjects with lower difficulties among the medium relevant subjects, as can be seen in Figure 2.11B.

Table 2.2: Pearson linear correlation coefficient between estimated difficulties and eigenvector centrality measures; and between estimated difficulties and strength centrality measures.

Group of subjects	Eigenvector		Strength		Number of subjects
	Correlation	p-value	Correlation	p-value	
all subjects	-0.62	0.0015	-0.57	0.0048	23
Low relevant subjects	-0.59	0.1589	-0.51	0.2382	7
Medium relevant subjects	-0.87	0.0006	-0.84	0.0013	11
High relevant subjects	0.16	0.796	-0.15	0.8142	5

It is worth mentioning that the difficulty scores for the subjects were estimated independently, without fitting an IRT model. Results support the claim that centrality measures, using the proposed multilayer network approach, are statistically correlated to item difficulty.

Results: The Brazilian Lottery Data Set

Figure 2.12A shows the overlapping network of the Brazilian lottery data set. The critical upper bound value was estimated as $\Psi = 21.07$, using a statistical significance level of 5%. Thus, matrix $\mathbf{U}$ has two connected lottery numbers (or vertices) if the same pair of numbers occurred, at least, in 22 games. Figure 2.12B shows the final network related to matrix $\mathbf{U}$.

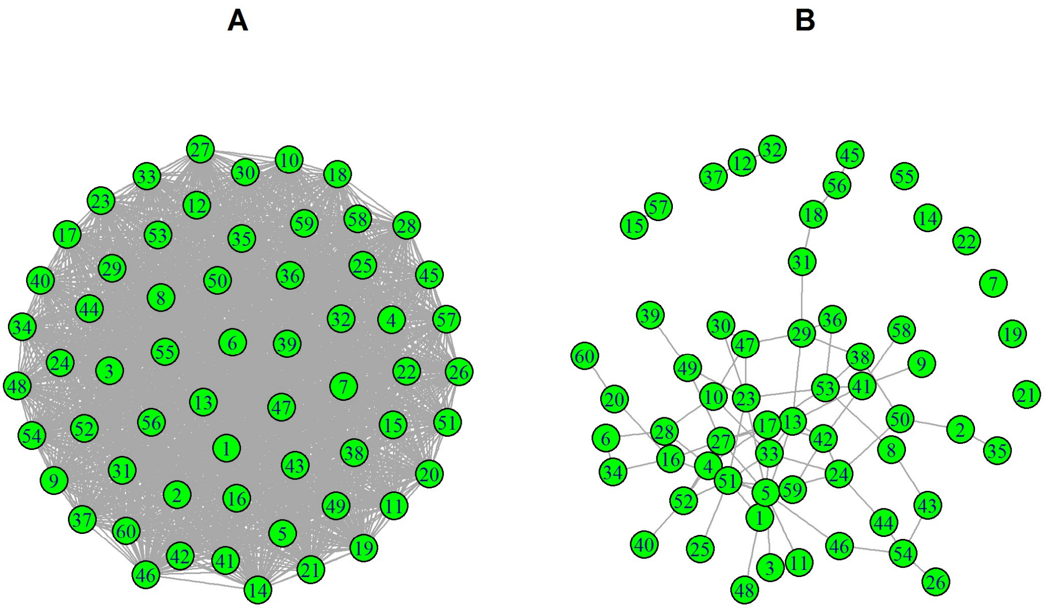

Figure 2.12: The Brazilian lottery network. (A) Single-layer network using the aggregation matrix $\mathbf{A}$. (B) Single-layer network using the aggregation matrix $\mathbf{U}$.

The p-value, estimated using equation 2.16, is 0.5308. Thus, there is statistical evidence that the null hypothesis of an independent multilayer network is true. That is, there is no pattern of numbers which are repeatedly being selected.

Discussion and Conclusion

This work proposes a network representation for questionnaire data in order to estimate network statistics which are correlated to item difficulty. In particular, we address the scenario in which the examinees are allowed to choose fewer items in the questionnaire. In this situation, IRT models fail to estimate the difficulties of the items. The proposed approach creates a multilayer network using the questionnaire data. In sequence, an overlapping network, identified using matrix **O**, is created using the multilayer network. Statistical procedures simplify the overlapping network generating a sparser network, known as matrix **U**. Centrality measures were investigated using both matrices **O** and **U**.

Results using simulated data and real case data sets show that centrality measures are strongly and statistically correlated to item difficulty. In the first and second simulated scenarios, the strength and the eigenvector centrality measures, using matrix **O**, are the most correlated measures to item difficulty. It is worth mentioning that the strength centrality measure using matrix **O** is proportional to the frequency of selected items. In the third scenario, the eigenvector centrality measure, using matrix **U,** is the most correlated measure to item difficulty. Nonetheless, our findings strongly suggest that the eigenvector centrality measure is the most consistent and robust statistic to estimate item difficulty, as compared to the remaining measures.

Furthermore, the proposed simplified network can be used to create a visual representation of the questionnaire, and the pairs of items which are the most frequently chosen by the examinees. This is particularly useful for educational researchers. This is illustrated using the Brazilian School of Engineering data set. It is worth noticing that the simplified network may have edges between items, even though the examinees are randomly choosing items. The proposed Monte Carlo statistical inference procedure provides a global statistical test which identifies whether the questionnaire multilayer network is independent.

Future work aims to propose IRT models using the eigenvector centrality measure in order to reduce bias estimates of the item difficulty. Two alternatives are currently being investigated: weighted log-likelihood maximization, using weights which are proportional to the eigenvector centrality measure; and, linear predictors of the item difficulty, using centrality measures as independent variables.

Acknowledgments

The authors thank Professor Walmir Matos Caminhas for valuable suggestions and for the school of engineering data set.

Chapter 3

A New Item Response Theory Model to Adjust Data Allowing Examinee Choice

Abstract

In a typical questionnaire testing situation, the examinees are not allowed to choose which items they would rather answer. The main reason is a technical issue in obtaining satisfactory statistical estimates of examinees' abilities and items' difficulties. This paper introduces a new Item Response Theory (IRT) model that incorporates information from a novel representation of the questionnaire data, using network analysis. Three scenarios in which examinees are allowed to select a subset of items were simulated. In the first scenario, the assumptions required to apply the standard Rasch model are met, thus establishing a reference of the parameters' accuracy. The second and third scenarios include five increasing levels of violation of those assumptions. Results show substantial improvements in the parameters' recovery over the standard model. Furthermore, the accuracy was closer to the reference in almost every evaluated situation. To the best of our knowledge, this is the first proposal to obtaining satisfactory IRT statistical estimates in these two last scenarios.

Keywords: Bayesian Modeling, Examinee choice, Item Response Theory, Missing data, Network analysis, Selection of items.

Introduction

Item response theory (IRT) comprises a set of statistical models for measuring examinees' abilities through their answers to a set of items (questionnaire). One of the most important advantages of IRT is to allow the comparison between examinees who answered to different tests. This property, known as invariance, is obtained by introducing separate parameters for the examinees' abilities and items' difficulties [Hambleton, Swaminathan and Rogers, 1991, p. 2 -8]. The IRT models have optimal properties when all items within a test are mandatory to the examinees. On the contrary, if the examinees are allowed to choose subset of items, instead of answer them all, the model estimates may became seriously biased. This problem has been raised by many researchers [Wainer, Wang and Thissen, 1994; Fitzpatrick and Yen, 1995; Wang, Wainer and Thissen, 1995; Bradlow and Thomas, 1998; Linn, Betebenner and Wheeler, 1998; Cankar, 2010; Wang et al., 2012], but still remains without a satisfactory proposal.

This is an important issue because several studies have provided evidences that choice has a positive impact in terms of educational development [Brigham, 1979; Baldwin, Magjuka and Loher, 1991; Cordova and Lepper, 1996; Siegler and Lemaire, 1997]. That is, they indicated that allowing students to choose increase the motivation and the depth of engagement in the learning process. In a testing situation, allowing choice seems to reduce the concern of examinees regarding the impact of an unfavorable topic [Jennings et al., 1999]. Besides, it has been claimed as a necessary step for the improvement of educational assessment [Rocklin, O'Donnell and Holst, 1995; Powers and Bennett, 2000].

Furthermore, it could be used in benefit of important challenges faced in the application of IRT models. For example, to achieve the invariance property the items used in different tests must be calibrated in the same scale. This is usually done by creating a bank of items from which the items used in all these tests shall be extracted. Items in the bank were previously calibrated by been exposed to examinees with similar features to whom the tests are intended. Therefore, these items were pre-tested. The pre-test process is typically extremely expensive and time-consuming. For example, in 2010 was reported that the Brazilian government spent about \$3.1 million to calibrate items for an important national exam (ENEM). Nevertheless, serious problems were reported to had occurred during the pre-test, like the supposedly illegal copy of many items by employees of a private high school college and their subsequent leakage. In addition, it was released that the number of items currently available in the national bank was about 6 thousand, whereas the ideal number would be between 15 and 20 thousand. All these events were harshly criticized by the mainstream media [Agência Brasil, 2010; Moura, 2010; Borges, 2011; S.P. and Mandelli, 2011]. Many recent researches have developed optimal designs for items calibration in order to reduce costs and time [Van der Linden and Ren, 2014; Lu, 2014]. Still, there is a limit for the number of items that an examinee can proper answer within a period of time. If the examinees are allowed to choose v items within a total o V items ($v < V$) and satisfactory statistical estimates are provided for all the V items, the costs of calibration per item are reduced.

This paper presents a new IRT model to adjust data generated using an examinee choice allowed scenario. The proposed model incorporates, in the IRT model, network analysis information using Bayesian modeling. Results show substantial improvements in the accuracy of the estimated parameters as compared to the standard IRT model, mainly in situations in which the examinees' choice are not a random selection of items. To the best of our knowledge, this is so far the only proposal that achieved a satisfactory parameters' estimation in some scenarios reported in literature, known as very critical scenarios to the standard Rasch model estimates [Bradlow and Thomas, 1998].

Material and Methods

The Standard IRT Model

The Rasch model [Rasch, 1960] is a widely used IRT model. The item characteristic curve is given by equation 3.1 [Hambleton, Swaminathan and Rogers, 1991, p. 12]:

$$P(Y_{i\alpha} = 1|\theta_\alpha, b_i) \ = \frac{e^{\theta_\alpha - b_i}}{1+e^{\theta_\alpha - b_i}} \ , \tag{3.1}$$

where $\mathrm{Y}_{\mathrm{i}\alpha} = \{0,1\}$ is a binary variable that indicates whether examinee α correctly answered item i; θ_α is the ability parameter of examinee α; b_i is the difficulty parameter of item i; $P(Y_{i\alpha} = 1|\theta_\alpha, b_i)$ is the probability that a randomly chosen examinee with ability θ_α correctly answers item i and the probability of an incorrect response is equal to $P(Y_{i\alpha} = 0|\theta_\alpha, b_i) = 1 - P(Y_{i\alpha} = 1|\theta_\alpha, b_i)$.

In the Rasch model, the b_i parameter represents the ability required for any examinee to have a 50% (0.50) chance of correctly answering item i. Given M examinees and V items, estimates for θ_α and b_i in the IRT models are found using the likelihood showed in equation 3.2 [Hambleton, Swaminathan and Rogers, 1991, p. 41]:

$$L(\vec{Y}|\vec{\theta}, \vec{b}) = \prod_{\alpha=1}^{M} \prod_{i=1}^{V} (P(Y_{i\alpha} = 1|\theta_\alpha, b_i))^{y_{i\alpha}} (1 - P(Y_{i\alpha} = 1|b_i, \theta_\alpha))^{1-y_{i\alpha}}, \tag{3.2}$$

where $\vec{\theta} = (\theta_1, \theta_2, \dots, \theta_M)$ is the vector of the examinees' abilities, $\vec{b} = (b_1, b_2, \dots, b_V)$ is the vector of the items' difficulties and $\vec{Y}$ is the vector of the observed responses. To use Bayesian estimation, the prior distributions $f(\theta_\alpha)$ and $f(b_i)$ must be defined. Since $\vec{\theta} \perp \vec{b}$, $\vec{\theta}_r \perp \vec{\theta}_s$, $\vec{b}_p \perp \vec{b}_q$, for $\mathrm{r} \neq \mathrm{s}$ and $\mathrm{p} \neq \mathrm{q}$, then the joint posterior distribution for parameters $\vec{\theta}$ and $\vec{b}$ is given by equation 3.3.

$$f(\vec{\theta}, \vec{b}|\vec{y}) \propto L(\vec{y}|\vec{\theta}, \vec{b}) \prod_{i=1}^{V} f(b_i) \prod_{\alpha=1}^{M} f(\theta_\alpha). \tag{3.3}$$

In the Rasch model, each latent ability θ_α is assumed to come from a standard normal distribution. Since the IRT model has more parameters than observations, this restriction is necessary. Further information regarding Bayesian estimation of IRT models is found in Fox

[2010]. Furthermore, a usual prior distribution for the item parameter is given by equation 3.4 [Albert, 1992; Patz and Junker, 1999; Curtis, 2010, Fox, 2010, p. 21]:

$$b_i|\mu_b, \sigma_b^2 \sim N(\mu_b, \sigma_b^2). \tag{3.4}$$

where $N(\mu_b, \sigma_b^2)$ denotes a normal distribution with mean of μ_b and variance of σ_b^2. One can use Markov Chain Monte Carlo (MCMC) methods to sample from the posterior distribution of μ_b and σ_b^2. Details about MCMC in the context of IRT are found in Patz and Junker [1999]. In the Appendix, a BUGS [Gilks, Thomas and Spiegelhalter, 1994] code for the adjustment of the standard Rasch model is found.

Inference under Examinee Choice Design

When allowing examinee choice, the likelihood equation includes the missing-data-indicator vector $\vec{R} = (R_{1,1}, R_{1,2}, \ldots, R_{i,\alpha}, \ldots, R_{V,\mathrm{M}})$, where $R_{i\alpha} = 1$ if examinee α response for item i is observed; otherwise, $R_{i\alpha} = 0$. Given $\vec{R}$, the set of responses $\vec{Y}$ can be written as $\vec{Y} = (\vec{\mathrm{Y}}_{obs}, \vec{\mathrm{Y}}_{mis})$ [Rubin, 1976], where $\vec{\mathrm{Y}}_{obs}$ denotes the observed values and $\vec{\mathrm{Y}}_{mis}$ denotes the missing values. Thus, the likelihood function given in equation 3.2 is rewritten [Bradlow and Thomas, 1998] as shown in equation 3.5:

$$L[(\vec{\mathrm{Y}}_{obs}, \vec{\mathrm{Y}}_{mis}), \vec{R}|\vec{\theta}, \vec{b}] = f[\vec{R}|(\vec{\mathrm{Y}}_{obs}, \vec{\mathrm{Y}}_{mis}), \vec{\theta}, \vec{b}]f[(\vec{\mathrm{Y}}_{obs}, \vec{\mathrm{Y}}_{mis})|\vec{\theta}, \vec{b}], \tag{3.5}$$

and the joint posterior distribution is given by equation 3.6:

$$f(\vec{\theta}, \vec{b}, \vec{Y}_{miss}|\vec{R}, \vec{Y}_{obs}) \propto f[\vec{R}|\vec{\mathrm{Y}}_{obs}, \vec{\mathrm{Y}}_{mis}, \vec{\theta}, \vec{b}]f[\vec{\mathrm{Y}}_{obs}, \vec{\mathrm{Y}}_{mis}|\vec{\theta}, \vec{b}] \prod_{i=1}^{V} f(b_i) \prod_{\alpha=1}^{M} f(\theta_\alpha). \tag{3.6}$$

Bradlow and Thomas [1998] showed that if examinees are allowed to choose items then valid statistical inference for $\vec{\theta}$ and $\vec{b}$ can be obtained using equation 3.2 only if the assumptions in equations (3.7) and (3.8) hold:

$$f\left[\vec{R}\middle|\left(\vec{\mathrm{Y}}_{obs}\,,\ \vec{\mathrm{Y}}_{mis}\right),\vec{\theta},\vec{b}\right] = f\left[\vec{R}\middle|\vec{\mathrm{Y}}_{obs},\vec{\theta},\vec{b}\right], \tag{3.7}$$

$$f\left[\vec{R}\middle|\vec{\mathrm{Y}}_{obs},\vec{\theta},\vec{b}\right] = f\left[\vec{R}\middle|\vec{\mathrm{Y}}_{obs}\right]. \tag{3.8}$$

Assumption (3.7) is known as the missing at random (MAR) assumption and implies that examinees are not able to distinguish which items they probably would answer correctly. Assumption (3.8) implies that examinees of different abilities generally do not broadly select the easier or the more difficult items. Further details are found in Bradlow and Thomas [1998]. If both assumptions (3.7) and (3.8) hold, then the posterior distribution can be rewritten as shown in equation 3.9:

$$f\left(\vec{\theta},\vec{b},\vec{Y}_{miss}\middle|\vec{R},\vec{Y}_{obs}\right) \propto f\left(\vec{Y}_{obs}\middle|\vec{\theta},\vec{b}\right) f\left(\vec{Y}_{miss}\middle|\vec{\theta},\vec{b}\right)\ \prod_{i=1}^{V} f(b_i) \prod_{\alpha=1}^{M} f(\theta_\alpha). \tag{3.9}$$

In this case, it is assumed that the process that generates missing data is non-informative. Details about statistical inference in the presence of missing data are found in Rubin [1976]. Since $\vec{\mathrm{Y}}_{\mathrm{miss}}$ is unknown, MCMC methods can be used to draw samples from the posterior distribution [Patz and Junker, 1999]. That is possible because equation 3.5 is an augmented data likelihood. Hereafter, it is assumed that if the examinees are randomly selecting items then the unobserved values are MAR; otherwise, the process that generates missing data is informative and the unobserved values are not MAR.

Wang, Wainer and Thissen [1995] conducted an experiment called "Choose one, Answer all" to test whether the MAR assumption is empirically plausible. In the experiment, 225 students indicated, among pairs of items, which items they would rather answer. Nevertheless, all items needed to be answered. Results showed that the MAR assumption did not hold. For example, a particular pair of items (items 11 and 12) was introduced to the students. One item (item 12) was very difficult as compared to the other item (item 11). It was observed that only 20% of the examinees preferred to answer item 12. It was further observed that those students who had chosen item 12 performed far better on item 11, which indicated that they had made a disadvantaged choice. Moreover, the examinees who chose item 11 performed better on both items as compared to those who chose item 12. These results were observed elsewhere (Chi, 1978; Chi, Glaser and Rees, 1982), suggesting that students with higher abilities are more able to differentiate difficulties between items.

Furthermore, Bradlow and Thomas [1998] performed simulation studies to demonstrate the bias in the estimated parameter, using the standard Rasch model in violation of assumptions (3.7) and (3.8). The complete data set had 5.000 examinees and 200 items. In the simulation study, 50 items were mandatory and the remaining 150 items were divided into 75 choice pairs. In the experiment in which the first assumption (3.7) was violated, there occurred a consistent underestimation of item difficulty for the 50 mandatory items and more severe underestimation for the remaining 75 choice pairs. Furthermore, in the experiment in which the second assumption (3.8) was violated, there occurred overestimation of item difficulty for high-difficulty items and underestimation for low-difficulty items. The authors also stated that very little is known about the nature and magnitude of realistic violations of those assumptions.

Wang et al. [2012] proposed the inclusion of a random-effect parameter γ_α in the standard IRT models to account for the choice effect. The proposed model produced better results in some simulated scenarios as compared to the standard IRT model. Nevertheless, the authors state that if the first (3.7) or the second assumption (3.8) is violated, as described in Bradlow and Thomas [1998], valid statistical inferences are not obtained for the standard IRT model, or for the proposed model with the random-effect parameter.

Network Information

In the Chapter 2 of this thesis, several simulation studies in violation of the MAR assumption were performed, in which M examinees choose V_α items from a total of V items,

$0 < V_\alpha \leq V$. A novel representation of examinees and their selected items using network analysis was proposed. The data set was coded as layers, vertices (or nodes) and edges. Briefly, a network (or graph) $\mathrm{G} = (\vec{\mathrm{V}}, \vec{\mathrm{E}})$ consists of a set of V vertices ($\vec{\mathrm{V}}$) that identify elements of a system and a set of E edges ($\vec{\mathrm{E}}$) that connect pairs of vertices $\{v_i, v_j\}$, pointing out the presence of a relationship between the vertices [Barrat, Barthélemy and Vespignani, 2008]. In the proposed network representation, each examinee is represented as a single network in which the V items are the vertices and every pair of the V_α selected items is connected by an edge. That is, the data set is initially represented as M single-layer networks, or a multilayer network $\vec{G} = (G_1, G_2, \dots, G_M)$ [Battiston, Nicosia and Latora, 2014]. From the multilayer network, two matrices are created.

The first matrix, called overlapping matrix, $\mathbf{O} = [o_{ij}]$, is a weighted $V \times V$ matrix in which the elements o_{ij} indicate the number of examinees that chose both items i and j [Bianconi, 2013; Battiston, Nicosia, Latora, 2014]:

$$o_{ij} = \sum_{\alpha=1}^{M} a_{ij}^{[\alpha]} \tag{3.10}$$

where $a_{ij}^{[\alpha]}$ = 1 if examinee α chose both items i and j and 0 otherwise, $0 \leq o_{ij} \leq M \ \forall\, i, j$. The second matrix, called matrix $\mathbf{U} = [u_{ij}]$, is a binary $V \times V$ matrix in which u_{ij} is equal to 1 if o_{ij} (equation 3.10) is greater than a threshold (Ψ) and zero otherwise. The threshold is calculated to identify recurrent edges in the multilayer network and is given by equation 3.11:

$$\Psi = \frac{E^{[o]}}{\eta_\varepsilon} + Z_\gamma \sqrt{\frac{E^{[o]}}{\eta_\varepsilon}(1 - \frac{1}{\eta_\varepsilon})}, \tag{3.11}$$

where $E^{[o]} = \sum_{\alpha=1}^{M} \frac{V_\alpha(V_\alpha - 1)}{2}$ is the total number of edges in the multilayer network, $\eta_\varepsilon = \frac{V(V-1)}{2}$ is the maximum number of edges in a single-layer network and Z_γ is a z-score statistic ($Z_{0,05} = 1.645$). Under the null hypothesis, the statistical distribution of the number of incident edges in each pair of vertices is the same. Thus, Ψ represents the upper bound of the observed number of edges between vertices i and j under the hypothesis that the $E^{[o]}$ edges are randomly distributed. Further details are found in the Chapter 2 of this thesis. Therefore, matrix **U** is a binary matrix that preserves only the statistically significant edges, as shown in equation 3.12:

$$u_{ij} = \begin{cases} 1, & if\ o_{ij} > \Psi;\ i \neq j; \\ 0, & otherwise. \end{cases} \tag{3.12}$$

In the Chapter 2 of this thesis, it was shown that the density of matrix **U** can be used to test whether the MAR assumption holds. Furthermore, the larger the density of matrix **U** the more violated the MAR assumption; that is, the density of matrix **U** indicates the violation level of the MAR assumption. The density of a network $\mathrm{G} = (\vec{\mathrm{V}}, \vec{\mathrm{E}})$ is given in equation 3.13 [Lewis, 2009, p. 53]:

$$\mathcal{D} = \frac{2E}{V(V-1)}. \tag{3.13}$$

Moreover, in the Chapter 2 several simulation studies using three different scenarios were performed. Several network centrality measures, and their correlations with item difficulty when MAR assumption is violated, were evaluated. Most frequently centrality measures found in literature [Batool and Niazi, 2014] were tested. The eigenvector of matrix **O** was found to be the most consistent and robust network statistic to estimate item difficulty. The eigenvector centrality of a vertex i (ρ_i) is the i-th element of the first eigenvector of matrix **O**:

$$\lambda\vec{\rho} = \mathbf{O}\vec{\rho} \tag{3.14}$$

where $\mathbf{O}$ represents the matrix **O**, λ is the eigenvalue and $\vec{\rho}$ is the first eigenvector of matrix **O**. Further details about eigenvector centrality are found in Bonacich [1972]. In addition, their simulation study indicates that the larger the MAR assumption violation, the larger the correlation between the eigenvector centrality and items difficulties. It is worth mentioning that the eigenvector centrality assumes values within the range 0–1. Therefore, it provides a standardized measure of vertex centrality.

The Proposed Model

In general, the relation between item difficulty b_i and first eigenvector of matrix **O** can be written as shown in equation 3.15.

$$b_i = g(\rho_i) \tag{3.15}$$

where $g(\rho_i)$ is a function of the i-th element of the first eigenvector $\vec{\rho}$. We propose a new IRT model that takes into account the relation shown in equation 3.15. This can be achieved defining the following prior distribution:

$$b_i|\mu_{b_i}, \sigma_b^2 \sim N(\mu_{b_i}, \sigma_b^2), \tag{3.16}$$

where $\mu_{b_i} = g(\rho_i)$ and σ_b^2 accounts for the variability of b_i that can not be explained by $g(\rho_i)$. It is worth mentioning that the larger the correlation between $\vec{\rho}$ and $\vec{b}$ the lower the dispersion parameter σ_b^2. The posterior distribution, shown in equation 3.6, can be rewritten as follows in equation 3.17:

$$\begin{aligned} f\left(\vec{\theta}, \vec{b}, \vec{Y}_{miss} \middle| \vec{\rho}, \vec{R}, \vec{Y}_{obs}\right) \propto f\left[\vec{R} \middle| \vec{\rho}, \vec{\mathrm{Y}}_{obs}, \vec{\mathrm{Y}}_{mis}, \vec{\theta}, \vec{b}\right] \\ \times f\left[\vec{\mathrm{Y}}_{obs}, \vec{\mathrm{Y}}_{mis} \middle| \vec{\rho}, \vec{\theta}, \vec{b}\right] \prod_{i=1}^{V} f(b_i\,|\rho_i)\ \prod_{\alpha=1}^{M} f(\theta_\alpha). \end{aligned} \tag{3.17}$$

If the assumptions (3.7) or (3.8) do not hold but, given $\vec{\rho}$, the equation 3.18 holds, the proposed model can provide valid statistical inference.

$$f\left[\vec{R} \middle| \vec{\rho}, \vec{\mathrm{Y}}_{obs}, \vec{\mathrm{Y}}_{mis}, \vec{\theta}, \vec{b}\right] \propto f\left[\vec{R} \middle| \vec{\rho}, \vec{\mathrm{Y}}_{obs}\right]. \tag{3.18}$$

Equation 3.18 assumes that, given $\vec{\rho}$, the missing-data-indicator $\vec{R}$ became independent of $\vec{\mathrm{Y}}_{mis}$, $\vec{\theta}$ and $\vec{b}$. Further information about conditioning on covariates for missingness mechanism become ignorable are found in [Little and Rubin, 1987, p. 9 -17; Bhaskaran and Smeeth, 2014].

In this paper, the following mathematical model between $\vec{\rho}$ and $\vec{b}$ is proposed:

$$b_i = -\frac{1}{\beta_1}\left(\log\left(\frac{1-(\rho_i-\beta_2)}{(\rho_i+\beta_2)-C}\right)+\beta_0\right) \tag{3.19}$$

where β_0, β_1 and β_2 are coefficients, to be estimated, and C is the minimum value of ρ_i, i. e., $C = \min(\vec{\rho})$. This model represents the inverse equation of a logistic function with a changed shape that asymptotically tends to the lowest value of $\vec{\rho}$ .The β_2 parameter is required to move the vector $\vec{\rho}$ bellow 1. Furthermore, the β_2 parameter is also used to shift vector $\vec{\rho}$ above its lowest value to prevent that $\log\left(\frac{1-(\rho_i-\beta_2)}{0}\right) = +\infty$. This model was empirically proposed based on the simulation studies, shown in the Results section. It is worth mentioning that other functions to describe the relation between $\vec{\rho}$ and $\vec{b}$ can be proposed. The BUGS code for the Proposed Model is available in the Appendix.

Simulation Study

To investigate the data behavior under the violations of assumptions (3.7) and (3.8), we performed several simulations using three different scenarios. In all scenarios, 1.000 examinees ($\alpha = 1, \dots, 1{,}000$) choose 20 items within a total of 50 items ($i = 1, \dots, 50$). The examinee ability (θ_α) and item difficulty (b_i) were both generated from a standard normal density distribution and held fixed. The complete data set can be represented by a 1,000 versus 50 dimensional matrix of binary responses. For each examinee α and item i the probability of a correct answer was calculated using equation 3.1. This probability was used to generate the outcome $Y_{i\alpha}$ using a Bernoulli random number generator.

In the first scenario, hereafter named scenario 1, both assumptions (3.7) and (3.8) were valid. That is, the items were randomly selected by the examinees and consequently the process that causes missing data was non-informative. Therefore, this is the scenario in which valid statistical inference can be obtained using the standard Rasch model.

The second scenario, named scenario 2, is identical to the first simulation scenario presented in the Chapter 2. In this scenario, each examinee chooses items based on current values of θ_α and b_i. That is, for each examinee α, the items are divided into two groups: the first group comprises items which are easier as compared to the examinee ability, i.e., $b_i \leq \theta_\alpha$. This is the group in which the examinee has a probability higher than 0.50 to achieve a correct answer. The second group comprises items which are more difficult as compared to the examinee ability, i.e., $b_i > \theta_\alpha$. In this group, the examinee has a probability lower than 0.50 to answer the items correctly. A weight value (w_i) is assigned to the items in each group. For items in group 2, the weight value is $w_i^{[2]} = 1$; whereas, for items in group 1, the weight value varies from 1.5 to 30: $w_i^{[1]} \in \{1.5, 2, 5, 10, 30\}$. For example, if $w_i^{[1]} = 2$, then it can be said that the items in group 1 have twice the chance of being selected by the examinee as compared to the items in group 2. In this scenario, assumptions (3.7) and (3.8) are violated.

Finally, in the third scenario, named scenario 3, the examinee choice depends on the y_{mis}. That is, assumption (3.7) is violated. Similar to scenario 2, for each examinee α, the items are divided into two groups. Nonetheless, in scenario 3, the first group comprises items that were correctly answered by the examinee in the complete data set ($y_{i,\alpha} = 1$) and the second group contains the items in which the examinee failed ($y_{i,\alpha} = 0$). Likewise scenario 2, for items in group 2, the weight value is $w_i^{[2]} = 1$; whereas, for items in group 1, the weight value varies from 1.5 to 30: $w_i^{[1]} \in \{1.5, 2, 5, 10, 30\}$. This scenario is similar to the second simulation study described in Bradlow and Thomas [1998].

It is worth mentioning that the selected items were generated using a multinomial probability distribution. That is, the probability of examinee α selecting item i is:

$$p_{i\alpha} = \frac{w_{i\alpha}}{\sum_i w_{i\alpha}}. \tag{3.20}$$

In scenario 1, $w_{i\alpha} = 1 \; \forall \; i$.

Results

In the Results section first we present empirical evidence of the proposed function given in equation 3.19 to describe the relation between $\vec{b}$ and $\vec{\rho}$. Second, we define values for the variance parameter of the prior distribution (σ_b^2). Prior values of the σ_b^2 paramater improve statistical properties of the proposed model**.** Finally, several simulation studies are performed in different conditions to compare the accuracy of parameters recovery obtained using the standard Rasch model and using our proposed model.

Empirical Validation of the Proposed Model

To evaluate the performance of the proposed function (equation 3.19) to predict item difficulty, using the first eigenvector of matrix **O**, we performed 10.000 Monte Carlo simulations for scenarios 2 and 3 and calculated the Residual Sum of Squares (RSS), shown in equation 3.21.

$$RSS = \sum_{i=i}^{V} (b_i - \hat{b}_i)^2 \tag{3.21}$$

where $\hat{b}_i$ is the fitted value of b_i using equation 3.19. It is worth mentioning that the smaller the value of RSS, the better the fit of the model and the lower the bias of the estimates. The β_0, β_1 and β_2 coefficients were estimated using the least-squares method adapted to a nonlinear model in software R [Fox and Weisberg, 2010; R Core Team, 2015]. Table 3.1 shows the minimum, mean, maximum and standard deviation values of the RSS for different weight values used in scenarios 2 and 3. In both scenarios, the larger the weights, the lower the RSS. Thus, there is empirical evidence that the proposed function seems to provide a better fit in situations in which the MAR assumptions is more violated.

Table 3.1: Summary of Residual Sum of Squares

Weight	Scenario 2				Scenario 3			
	Min	Mean	Max	Sd	Min	Mean	Max	Sd
1.5	4.6521	10.0364	20.5177	1.8479	7.7194	16.793	32.3424	3.3918
2	2.1942	5.8299	11.0088	1.2245	2.9474	7.4835	13.6649	1.5131
5	0.6577	2.5018	5.5351	0.825	0.7303	1.9269	4.2557	0.4335
10	0.4425	1.9922	4.6373	0.7541	0.4524	1.1196	2.4593	0.2501
30	0.2682	1.8019	4.4424	0.7522	0.2799	0.738	1.5916	0.1545

Figure 3.1A shows the density of matrix **U** versus the RSS, using data generated from scenario 2. Figure 3.1B shows the density of matrix **U** versus the RSS, using data generated from scenario 3. In general, the larger the density of matrix **U**, the lower the RSS. In the Chapter 2 of this thesis, was shown that the larger the density of matrix **U**, the stronger the MAR assumption violation. Therefore, the density of matrix **U** can be used as a predictor of the goodness-of-fit statistic of the proposed model, shown in equation 3.19.

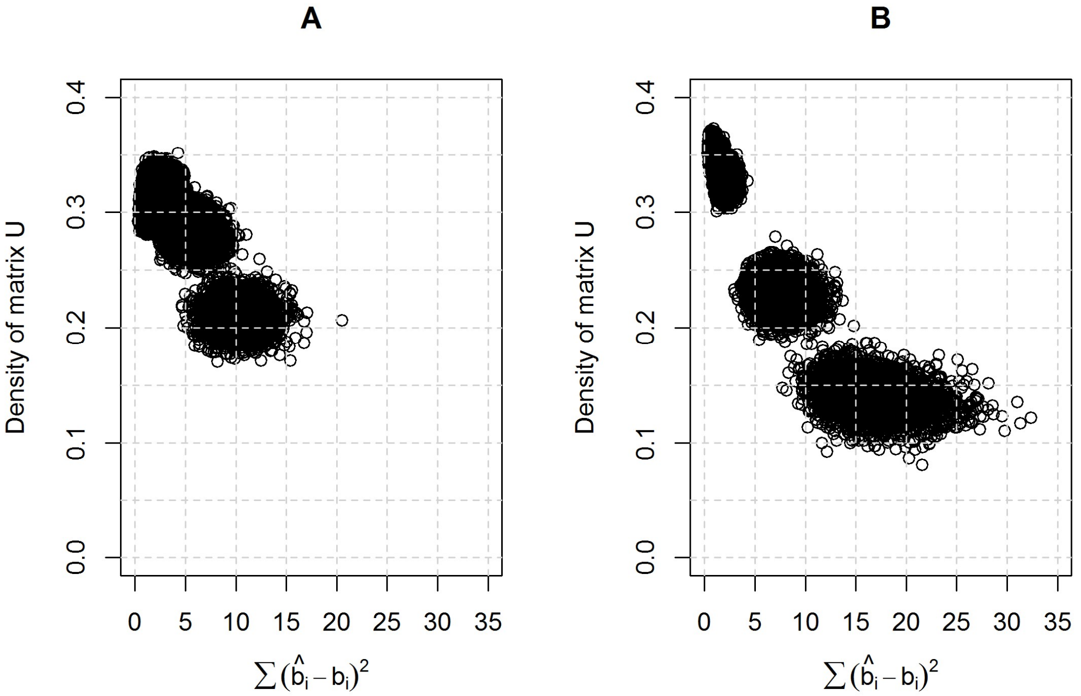

Figure 3.1: Residual Sum of Squares versus density of matrix U. (A) Scenario 2. (B) Scenario 3.

We propose to use the observed density value of matrix **U** to chose different values for σ_b^2 in the prior distribution (equation 3.16). Lower prior values for σ_b^2 mean that the posterior distribution of b_i will be concentrated towards its mean, $\mu_{b_i} = g(\rho_i)$, i.e., the posterior estimate of b_i is mostly defined by the proposed eigenvector centrality function. On the contrary, the larger the prior value of σ_b^2, the lower the posterior estimate of b_i is affected by the eigenvector centrality function.

Based on the results showed in Figure 3.1 and further simulation studies, the proposed values of σ_b^2 are showed in Table 3.2. As previously mentioned, the lower the density of matrix **U**, the larger the prior value of σ_b^2. Furthermore, if the density of matrix **U** is within 0.0–0.1, there is empirical evidence that the MAR assumptions holds. In this case, a large variance value is chosen in order to make the prior distribution for b_i less informative. Future work includes exhaustive simulation studies to propose a mathematical function that relates the density of matrix **U** and the prior value of σ_b^2, therefore, further minimizing the bias of the proposed model.

Table 3.2: The proposed values for the variance of the prior distribution

Density of matrix U	Proposed value for σ_b^2
[0 - 0.10]	10
(0.10 - 0.20]	5
(0.20 - 0.25]	0.5
(0.25 - 0.30]	0.4
(0.30 - 0.35]	0.3
>=0.35	0.2

Figure 3.2 illustrates the proposed function. It shows the plot of b_i versus ρ_i for 10 Monte Carlo simulations. Black circles represent data generated from scenario 1 ($w_i = 1$), in which there is no correlation between $\vec{b}$ and $\vec{\rho}$. With exception of the black circles, Figure 3.2A shows data generated from scenario 2 and Figure 3.2B data generated from scenario 3 using three different weight values, $w_i \in \{2, 5, 30\}$. The proposed function was adjusted using the mean value of ρ_i for each simulated weight, that is $\bar{\rho}_i = \sum_{i=1}^{10} \rho_i$, and is represented as the black lines. In both scenarios, the proposed function achieved a good data fit.

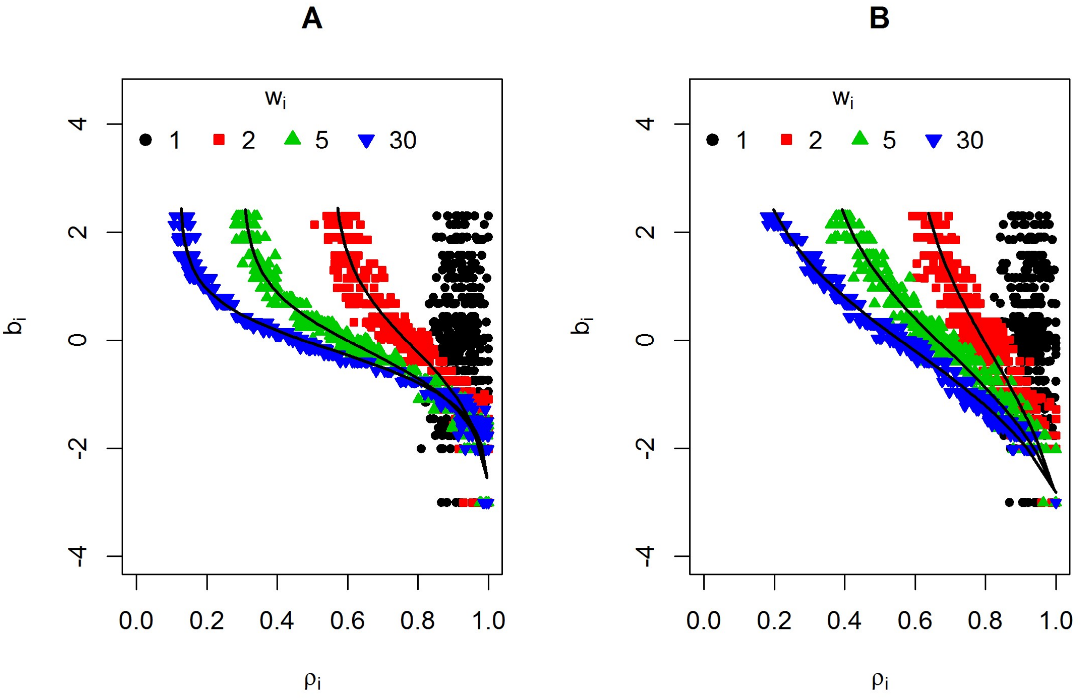

Figure 3.2: Item difficulty versus the first eigenvector of matrix O. (A) Scenario 2. (B) Scenario 3. Black lines represent the proposed function (equation 3.19).

Accuracy of the Estimated Parameters

In order to evaluate the performance of the proposed model as compared to the standard Rasch model, 100 Monte Carlo simulations were performed under the three scenarios, previously described. Scenario 1 is used as the reference scenario, since in this scenario valid statistical inferences can be obtained using the standard Rasch model [Bradlow and Thomas, 1998]. The accuracy of the estimated parameters is evaluated using the bias, the root mean square error (RMSE) and the root maximum square error ($\mathrm{RSE_{MAX}}$), given by equations (3.22), (3.23) and (3.24).

$$Bias(b_i) = \sum_{n=1}^{100}(\hat{b}_{ni} - b_i)/100 \qquad (3.22)$$

$$RMSE(b_i) = \sqrt{\frac{1}{100} \times \sum_{n=1}^{100}(\hat{b}_{ni} - b_i)^2} \qquad (3.23)$$

$$RSE_{MAX}\,(b_i) = \sqrt{Max\,(\cup_{n=1}^{100}\,\{(\hat{b}_{ni} - b_i)^2\}\,)} \tag{3.24}$$

where b_i is the true difficulty of item i, $\hat{b}_{ni}$ is the estimated value of b_i in the *n-th* Monte Carlo simulation. In addition, 95% highest probability density (HPD) intervals [Chen and Shao, 1999] are provided for each estimated difficulty. Using the HPD intervals, the proportion of simulated parameters that fall within the HPD intervals are calculated. It is worth mentioning that the expected a posterior (EAP), i.e., the mean of the posteriori distribution was used as the point estimate of b_i [Kieftenbeld and Natesan, 2012].

The standard Rasch model and the proposed model were estimated using the BUGs codes shown in Table 3.6 and Table 3.7, in Appendix. The WinBugs [Lunn et al., 2000] software was used. The MCMC simulations used 10.000 iterations, a burn-in period of 1.000 iterations and a sample period (thin) of 5 iterations. The variance of the prior distribution was selected as 10, for the standard Rasch model, and was selected according to Table 3.2 for the proposed model. Initial values of the parameters were: $\beta_0 = 0,\ \beta_1 = -1.5,\ \beta_2 = 0.01,\ \mu_b = 0,\ b_i = 0\ \forall i,\ \theta_\alpha = 0\ \forall\alpha$. The convergence of the model was evaluated using the autocorrelation level of the chains, which was near to zero, and the trace plot, which indicated that the convergence was achieved. The average time to run the standard Rasch model once was 15.10 minutes and the average time to run the proposed model was 15.40 minutes using a Intel (R) Core i7 processor with 2 GHz and 8 GB of RAM. Further details about model specifications in WinBugs can be found in Spiegelhalter et al. [2003].

Figure 3.3 shows differences between the estimates using the standard Rach model and the true items' difficulties, plotted against the true items' difficulties. Figure 3.3A shows the results using data generated from scenario 1. Figure 3.3B and 3.3C show the results using data generated from scenario 2 and scenario 3, respectively, using weight value equals to 10. Figure 3.3A shows that the residues $(\hat{b}_i - b_i)$ behave as a random cloud around zero. On the contrary, Figure 3.3B shows that the dispersion of the residues around zero are larger for more difficult items. Figure 3.3C shows a severe underestimation of the items' difficulties. Figure 3.4 shows the estimated 95% HPD intervals and the true items' difficulties (red points). The HPD obtained using the data set generated from scenario 1 (Figure 3.4A) and scenario 2 (Figure 3.4B) contains most of the true items' difficulties. It is worth noticing that some of the HPD intervals shown in Figure 3.4B are larger as compared to Figure 3.4A. On the contrary, all the HPD intervals shown in Figure 3.4C (scenario 3) do not contain the true item's difficulties, which shows a serious model fitting problem.

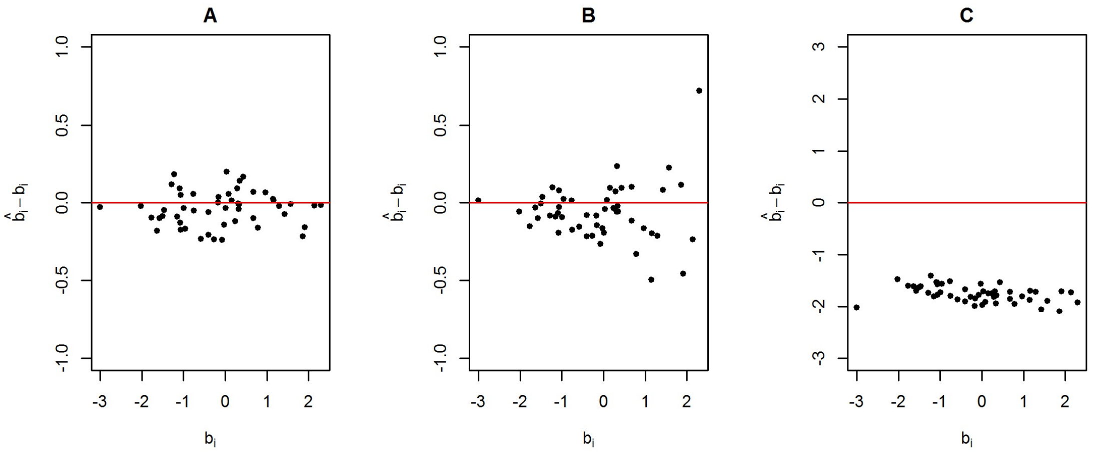

Figure 3.3: Estimates minus true items' difficulties plotted against true items' difficulties for three simulated data sets. (A) Data set generated from scenario 1. (B) Data set generated from scenario 2 using weight value equals to 10. (B) Data set generated from scenario 3 using weight value equals to 10.

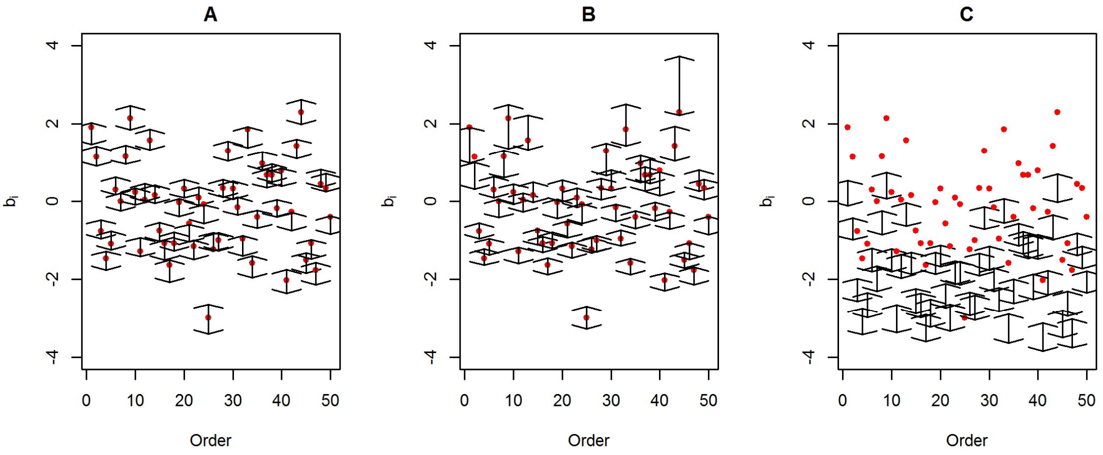

Figure 3.4: 95% HPD plotted according the item order for three simulated data sets. Red points represent the true items' difficulties. (A) Data set generated from scenario 1. (B) Data set generated from scenario 2 using weight value equal to 10. (B) Data set generated from scenario 3 using weight value equal to 10.

Empirical results showed that the fitting problem reported in scenario 3 can not be solved by replacing the standard Rasch model for the proposed model. It can be seen that the variability of the bias is lower, nevertheless, the estimated difficulties are consistently shifted toward lower values. We suggest fixing this problem by fitting the proposed model in two stages. In the first stage, the examinees are free to choose v within V items. In the second stage, a few of those V items will be set as mandatory (v_c) and the examinees are required to answer, again, these v_c items. Thus, these v_c items will be estimated twice as follows: in the first stage the v_c items are adjusted using our proposed method; in the second stage, the complete responses of the v_c items are adjusted separately using the standard IRT model. The average difference between the estimated difficulties in the second and the first stages of the v_c items is calculated and added to the first difficulty estimates of all V items. This procedure adds a constant value to the estimates obtained in the first stage, thus correcting the bias shift in scenario 3.

Two items were used to perform this calibration. Results show that two items were sufficient to correct the bias shift. Future work includes more exhaustive simulation studies to define the number of mandatory items in the second stage. It is worth mentioning that this procedure is not necessary in scenario 2 since the residues are already centered at zero. Nevertheless, in a real case situation, is it not possible to identify whether the choice was made following scenario 2 or scenario 3. Nonetheless, this procedure can be applied in both scenarios.

Table 3.3 shows parameters' accuracy for scenario 1 using the standard Rasch model and the proposed model. As previously mentioned, the standard Rasch model achieves valid statistical inference in this scenario and will be used as the accuracy reference of the estimated parameters. Thus, the reference model has minimum and the maximum Bias of -0.2542 and 0.1648, respectively. The mean and the maximum values for the RMSE are 0.1225 and 0.2814, respectively. The $\mathrm{RSE}_{\mathrm{MAX}}$ represents the largest absolute difference between the true and estimated difficulty for each item, the mean value is 0.3251 and the maximum value is 0.7342. That is, the largest difference between true and estimated items difficulties in scenario 1 is 0.7342. The width of the 95% HPD interval is 0.4996, on average, and the maximum width is 0.8198. Finally, on average, 95.30% of the true items' difficulties is inside the 95% HPD interval and the minimum proportion is 82%.

Table 3.3: Summary of estimated parameters in scenario 1

	STANDARD RASCH MODEL					PROPOSED MODEL				
	BIAS	RMSE	RSE MAX	HPD WIDTH	INSIDE HPD	BIAS	RMSE	RSE MAX	HPD WIDTH	INSIDE HPD
Min	-0.2542	0.0718	0.1732	0.4480	82.0%	-0.257	0.0835	0.2134	0.4509	86.0%
Mean	-0.0248	0.1225	0.3251	0.4996	95.3%	-0.0227	0.1231	0.3111	0.5014	95.0%
Max	0.1648	0.2814	0.7342	0.8198	100.0%	0.1699	0.2831	0.5715	0.8257	100.0%
Sd	0.0755	0.0359	0.1013	0.0671	3.44%	0.0784	0.0356	0.0704	0.0675	2.72%

Furthermore, Table 3.3 shows that the proposed model is suitable to fit data generated from scenario 1. This is because the variance of the prior distribution when the density of matrix **U** is close to zero is sufficiently large, which makes the prior distribution less informative. It is worth noticing that the maximum value of $\mathrm{RSE}_{\mathrm{MAX}}$ is lower in the proposed model (0.5715) as compared to the standard Rasch model (0.7342). This is because even though the data sets were generated using a random selection of items, a correlation between the eigenvector and the items difficulties can eventually occur in a few simulations. This may be the reason why the maximum value of $\mathrm{RSE}_{\mathrm{MAX}}$ is lower in the proposed model.

Table 3.4 shows accuracies of estimated parameters for scenario 2 using the standard Rasch model and the proposed model, with five levels of MAR assumption violation. The standard Rasch model results are close to scenario 1 for lower weight values: $w_i = 1.5$ and $w_i = 2$. For larger levels of MAR assumption violation, $w_i = 5$, $w_i = 10$ and $w_i = 30$, the Bias, RMSE and $\mathrm{RSE}_{\mathrm{MAX}}$ values are different from those observed in scenario 1. On the contrary, results using the proposed model are closer to those observed in scenario 1, even for larger weights. Therefore, our proposed model presents lower Bias, RMSE and $\mathrm{RSE}_{\mathrm{MAX}}$ values, narrower 95% HPD width and a larger average proportion of the true items' difficulties inside HPD intervals.

Table 3.4: Summary of parameters recovery in scenario 2

	STANDARD RASCH MODEL					PROPOSED MODEL				
	Weight = 1.5					Weight = 1.5				
	BIAS	RMSE	RSE MAX	WIDTH HPD	INSIDE HPD	BIAS	RMSE	RSE MAX	WIDTH HPD	INSIDE HPD
Min	-0.2407	0.0733	0.1857	0.4497	68.00	-0.2058	0.0733	0.1773	0.4431	66.00
Mean	-0.0223	0.1333	0.3599	0.5013	93.34	-0.0190	0.1297	0.3507	0.4898	93.66
Max	0.1528	0.2660	0.6305	0.7535	100.00	0.1507	0.2356	0.5860	0.6885	100.00
Sd	0.0750	0.0326	0.0833	0.0708	5.41	0.0720	0.0308	0.0774	0.0610	5.30
	Weight = 2					Weight = 2				
	BIAS	RMSE	RSE MAX	WIDTH HPD	INSIDE HPD	BIAS	RMSE	RSE MAX	WIDTH HPD	INSIDE HPD
Min	-0.2455	0.0809	0.1751	0.4453	74.00	-0.2028	0.0634	0.1692	0.4323	70.00
Mean	-0.0246	0.1387	0.3748	0.5054	92.22	-0.0255	0.1278	0.3273	0.4816	92.94
Max	0.1476	0.2693	0.7198	0.7734	100.00	0.1611	0.2286	0.5586	0.6848	100.00
Sd	0.0788	0.0350	0.1071	0.0830	5.12	0.0740	0.0335	0.0770	0.0644	5.08
	Weight = 5					Weight = 5				
	BIAS	RMSE	RSE MAX	WIDTH HPD	INSIDE HPD	BIAS	RMSE	RSE MAX	WIDTH HPD	INSIDE HPD
Min	-0.2168	0.0738	0.1754	0.4135	64.00	-0.1821	0.0717	0.1635	0.3963	66.00
Mean	-0.0311	0.1454	0.3941	0.5315	91.96	-0.0243	0.1224	0.3162	0.4825	94.70
Max	0.1818	0.3147	0.8963	1.0369	100.00	0.1461	0.2299	0.6164	0.7751	100.00
Sd	0.0824	0.0509	0.1536	0.1497	7.17	0.0690	0.0371	0.1000	0.1004	4.85
	Weight = 10					Weight = 10				
	BIAS	RMSE	RSE MAX	WIDTH HPD	INSIDE HPD	BIAS	RMSE	RSE MAX	WIDTH HPD	INSIDE HPD
Min	-0.2356	0.0757	0.2234	0.3970	64.00	-0.1848	0.0596	0.1602	0.3830	72.00
Mean	-0.0365	0.1467	0.3930	0.5594	92.88	-0.0297	0.1238	0.3228	0.4965	94.10
Max	0.2016	0.3485	1.1578	1.2580	100.00	0.1441	0.2544	0.6915	0.8789	100.00
Sd	0.0885	0.0636	0.1718	0.2116	6.56	0.0739	0.0481	0.1262	0.1327	4.58
	Weight = 30					Weight = 30				
	BIAS	RMSE	RSE MAX	WIDTH HPD	INSIDE HPD	BIAS	RMSE	RSE MAX	WIDTH HPD	INSIDE HPD
Min	-0.3008	0.0654	0.1551	0.3837	64.00	-0.2548	0.0591	0.1234	0.3715	72.00
Mean	-0.0321	0.1664	0.4247	0.6061	91.74	-0.0479	0.1365	0.3321	0.5260	93.44
Max	0.3197	0.5310	1.3339	1.7749	100.00	0.1224	0.3141	0.6980	1.0569	100.00
Sd	0.1213	0.1009	0.2573	0.3192	6.21	0.0878	0.0632	0.1335	0.1843	5.20

Table 3.5 shows results using scenario 3. The RSME and $\mathrm{RSE}_{\mathrm{MAX}}$ values for both models are larger than those values observed in scenario 2. Results of the proposed model are different from those values observed in scenario 1 only for $w_i = 30$. Nevertheless, the proposed model presents better results as compared to the standard Rasch model, especially for larger levels of MAR assumption violation ($w_i = 5, w_i = 10, w_i = 30$). In general, main conclusions are similar to those found for scenario 2.

In addiction, it is worth noticing that the largest difference between true and estimated items' difficulties is 0.7082 in the proposed model, whereas using the standard Rasch model the largest difference is 1.3339. This improvement is significant since the scale of the items' difficulties is based on a standard normal distribution.

Table 3.5: Summary of parameters recovery in scenario 3

STANDARD RASCH MODEL						PROPOSED MODEL				
	Weight = 1.5					Weight = 1.5				
	BIAS	RMSE	RSE MAX	WIDTH HPD	INSIDE HPD	BIAS	RMSE	RSE MAX	WIDTH HPD	INSIDE HPD
Min	-0.2444	0.0811	0.1954	0.4538	68.00	-0.2370	0.0806	0.1798	0.4560	76.00
Mean	-0.0254	0.1386	0.3778	0.5066	92.70	-0.0328	0.1373	0.3640	0.5085	92.86
Max	0.1593	0.2778	0.6631	0.8751	100.00	0.1493	0.2700	0.6367	0.8722	100.00
Sd	0.0749	0.0349	0.0889	0.0717	5.59	0.0763	0.0340	0.0892	0.0718	4.72
	Weight = 2					Weight = 2				
	BIAS	RMSE	RSE MAX	WIDTH HPD	INSIDE HPD	BIAS	RMSE	RSE MAX	WIDTH HPD	INSIDE HPD
Min	-0.2286	0.0823	0.1937	0.4631	74.00	-0.2263	0.0868	0.2217	0.4522	64.00
Mean	-0.0294	0.1420	0.3764	0.5159	92.88	-0.0469	0.1412	0.3600	0.5000	91.70
Max	0.1410	0.2841	0.6047	0.8906	100.00	0.1291	0.2507	0.5056	0.7951	100.00
Sd	0.0717	0.0320	0.0870	0.0730	5.93	0.0723	0.0311	0.0728	0.0608	7.05
	Weight = 5					Weight = 5				
	BIAS	RMSE	RSE MAX	WIDTH HPD	INSIDE HPD	BIAS	RMSE	RSE MAX	WIDTH HPD	INSIDE HPD
Min	-0.2104	0.0731	0.2066	0.5098	66.00	-0.2104	0.0615	0.1343	0.4736	72.00
Mean	-0.0565	0.1611	0.4364	0.5743	91.98	-0.0461	0.1339	0.3452	0.5205	93.96
Max	0.1110	0.2780	0.7341	1.0585	100.00	0.1111	0.2354	0.4734	0.8205	100.00
Sd	0.0691	0.0375	0.1119	0.0913	6.61	0.0702	0.0333	0.0734	0.0607	5.38
	Weight = 10					Weight = 10				
	BIAS	RMSE	RSE MAX	WIDTH HPD	INSIDE HPD	BIAS	RMSE	RSE MAX	WIDTH HPD	INSIDE HPD
Min	-0.2967	0.0907	0.2690	0.5681	54.00	-0.2572	0.0559	0.1573	0.5050	64.00
Mean	-0.0847	0.1945	0.5134	0.6416	89.56	-0.0742	0.1485	0.3615	0.5511	92.40
Max	0.0970	0.3312	1.0218	1.2108	100.00	0.1193	0.2759	0.5444	0.8494	100.00
Sd	0.0809	0.0453	0.1375	0.1051	8.97	0.0819	0.0431	0.0788	0.0598	7.66
	Weight = 30					Weight = 30				
	BIAS	RMSE	RSE MAX	WIDTH HPD	INSIDE HPD	BIAS	RMSE	RSE MAX	WIDTH HPD	INSIDE HPD
Min	-0.4101	0.1066	0.2877	0.6803	56.00	-0.3770	0.0623	0.1640	0.5362	62.00
Mean	-0.0970	0.2436	0.6307	0.7613	86.74	-0.0684	0.1710	0.3785	0.5757	87.20
Max	0.1275	0.4483	1.0165	1.4042	98.00	0.3421	0.3887	0.7082	0.8335	100.00
Sd	0.1260	0.0639	0.1592	0.1164	9.09	0.1428	0.0787	0.1132	0.0496	7.22

Figure 3.5 shows differences between the proposed model and the standard Rasch model using a data set generated from scenario 2, with weight equal to 10. The same data set was used in Figure 3.3B. Figure 3.5A presents the residues of the standard Rasch model and

Figure 3.5B presents the residues of the proposed model. Results shown in Figure 3.5B seems to be a random cloud centered at zero, as observed in Figure 3.3A. Furthermore, differences between the true and estimated items' difficulties are, in general, lower for the proposed model.

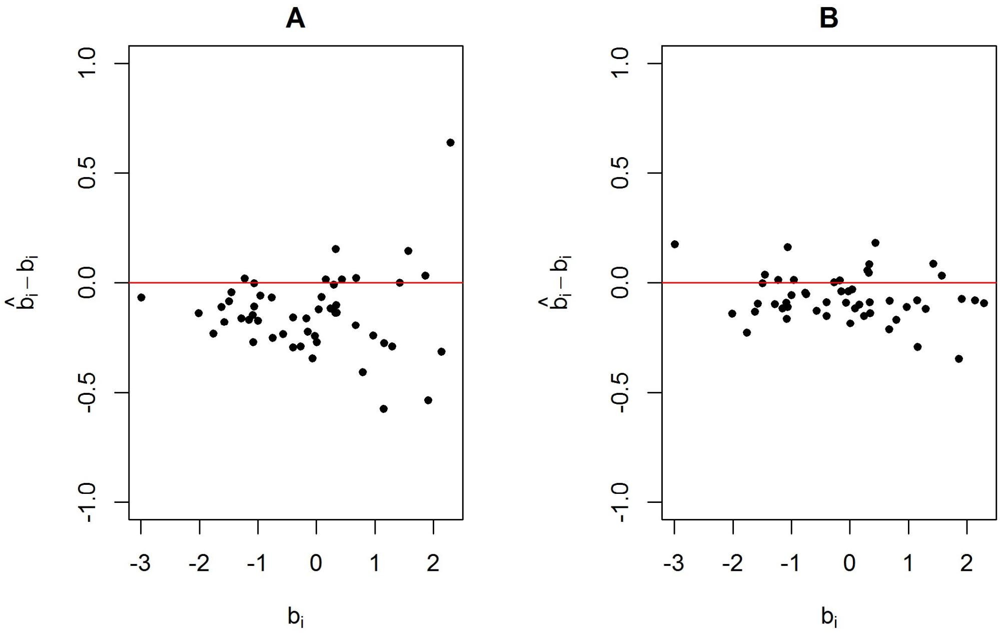

Figure 3.5: Differences between estimated and true items' difficulties plotted against true items difficulties, using data set generated from scenario 2, with weight value equals to 10. (A) Estimated values using the standard Rasch model. (B) Estimated values obtained using the proposed Model.

Figure 3.6 shows differences between the proposed model and the standard Rasch model using the same data presented in Figure 3.3C. The proposed two stage procedure is able to center the residues at zero. That is, the severe underestimation problem observed in Figure 3.3.3C is fixed. Figure 3.6A shows the results of the standard Rasch model. It is worth noticing that these results are the previously estimated results, shown in Figure 3.3C, with the addition of the constant shift. Figure 3.6B shows results using the proposed model. The residues of the proposed model are closer to a random cloud centered at zero, as observed in Figure 3.3A, compared to residues of the standard Rasch model.

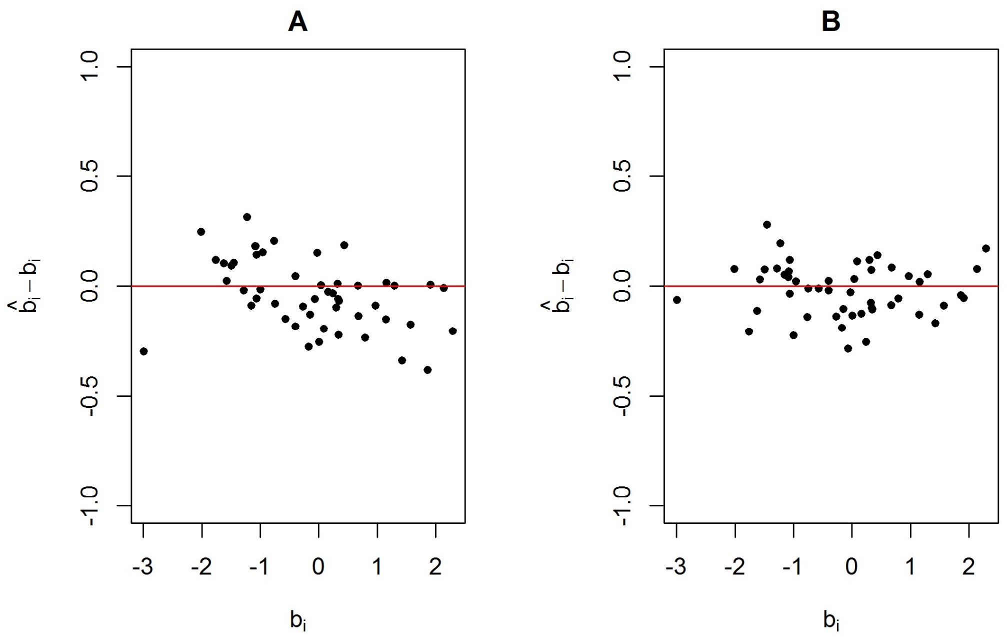

Figure 3.6: Differences between estimated and true items' difficulties plotted against true items difficulties, using data set generated from scenario 3, with weight value equals to 10. (A) Estimated values using the standard Rasch model. (B) Estimated values using the proposed model.

Discussion and Conclusion

This paper presents a new IRT model to estimate items' difficulties using questionnaire data in which examinees are allowed to choose a subset of items. Using network analysis, new information is incorporated in the model. Results are presented using three simulation scenarios. In the first scenario, the assumptions required in literature to apply the standard Rasch model are met. This scenario is used as the reference scenario. Second and third scenarios include five increasing levels of violations of those assumptions. These scenarios are reported in IRT literature and, to the best of our knowledge, none existing proposal had provided satisfactory results.

In both scenarios 2 and 3, the standard Rasch model is robust for lower levels of violations. The proposed model performs better in both scenarios, especially for larger levels of violations. Furthermore, the proposed model is able to achieve accuracies of the estimated parameters closer to reference values for both scenarios and different levels of violations, except for the largest violation level in scenario 3. As our main conclusion, we strongly recommend the use of the proposed model over the standard Rasch model, if allowing examinees to choose items.

In order to overcome further technical issues that prevents allowing examinee choice in practical situations, further investigations are required. For instance, real data set analyses are a crucial point to understand how examinees actually make their choices. Experiments in which examinees are required to indicate which subset of items they would rather answer, despite of all items being mandatory, could be used. Furthermore, some important issues like the number of items and which items should be mandatory for the two stage adjustment were out of the scope of this paper, and requires further investigation. Finally, the proposed approach requires further evaluation and research in order to be used in more complex IRT models.

Appendix

Table 3.6 shows the BUGS code for the standard Rasch model estimation using MCMC method. This BUGs code can be run using one of the following free softwares: WinBugs [Lunn, et al., 2000], OpenBugs [Thomas et al., 2006] and JAGS [Plummer, 2003]. Generally speaking, the BUGs code comprises the likelihood function (lines 2-5) and the prior distributions (lines 6-9). It is worth mentioning that BUGs language uses the precision parameter as opposed to the variance parameter for the normal. Thus, line 10 shows the precision parameter as a function of the standard deviation. Further details about BUGS code to fit IRT models can be found in Curtis [2010].

Table 3.6: BUGS code for the standard Rasch Model

```
1 model <- function(){
2     for( alfa in 1 : M ) {
3           for(i in 1:V){
4                 y[alfa,i] ~ dbern(prob[alfa,i])
5                 logit(prob[alfa,i])<-theta[alfa]-b[i] }}
6     for ( alfa in 1 : M ) {
7          theta[alfa] ~ dnorm(0,1)}
8     for(i in 1:V){
9           b[i] ~ dnorm(m.b,pr.b)}
10    pr.b<-pow(s.b,-2)}
```

Table 3.7 shows the BUGS code for the proposed model estimation. Lines 2-5 comprise the likelihood function, lines 6-8 shows the hyperparameters' prior distributions (β_0, β_1 and β_2) and lines 9-14 shows the parameters' prior distributions (θ_α and b_i). It is worth noticing that Line 8 shows that the prior density distribution of β_2 was truncated in the left at 0.0001 in order to avoid division by zero in equation 3.19, that is, $\beta_2 > 0$. Furthermore, in Line 12, "rho[i]" represents ρ_i and, in Line 13, "C" represents the lowest value of $\vec{\rho}$; both values are data driven.

Table 3.7: BUGS code for the Proposed Model

```
model <- function(){
    for( alfa in 1 : M ) {
          for(i in 1:V){
               y[alfa,i] ~ dbern(prob[alfa,i])
               logit(prob[alfa,i])<-theta[alfa]-b[i] }}
    beta0~dnorm(0,0.001)
    beta1~dnorm(0, 0.001)
    beta2~dnorm(0, 0.001)I(0.0001,)
    for( alfa in 1 : M ) {
          theta[alfa] ~ dnorm(0, 1)}
    for(i in 1:V){
          m.b[i]<- -1/beta1 * (log((1-rho[i] + beta2)/(rho[i] + beta2 - C)) + beta0)
          b[i] ~ dnorm(m.b[i],pr.b[i]) }
    pr.b<-pow(s.b,-2)}
```

References

1. Agência Brasil. Custos para pré-testar itens do Enem podem chegar a R$6 milhões. O Globo. 2010. Available from: http://oglobo.globo.com/sociedade/educacao/custos-para-pre-testar-itens-do-enem-podem-chegar-r-6-milhoes-2965286
2. Albert JH. Bayesian estimation of normal ogive item response curves using Gibbs sampling. Journal of Educational Statistics. 1992; 17: 251-269.
3. Baldwin TT, Magjuka RJ, Loher BT. The perils of participation: effects of choice of training on trainee motivation and learning. Personnel Psychology. 1991; 44: 51–65.
4. Barigozzi M, Fagiolo G, Garlaschelli D. Multinetwork of international trade: A commodity-specific analysis. Physical Review E. 2010; 81(4): 046104.
5. Barrat A, Barthélemy M, Pastor-Satorras R, Vespignani A. The architecture of complex weighted networks. PNAS. 2004; 101: 3747–3752.
6. Barrat A, Barthélemy M, Vespignani A. Dynamical processes in complex networks. Cambridge University Press; 2008. ISBN 978-0-521-87950-7.
7. Barrett L, Henzi SP, Lusseau, D. Taking sociality seriously: the structure of multidimensional social networks as a source of information for individuals. Phil. Trans. R. Soc. B. 2012; 367:2108–2118.
8. Batool K, Niazi MA. Towards a methodology for validation of centrality measures in complex networks. PloS one. 2014; 9(4):e90283.
9. Battiston F, Nicosia V, Latora V. Structural measures for multiplex networks. Physical Review E. 2014; 89(3):032804.
10. Bhaskaran K, Smeeth L. What is the difference between missing completely at random and missing at random? International journal of epidemiology. 2014. ISSN 0300-5771 DOI: 10.1093/ije/dyu080.

11. Bianconi G. Statistical mechanics of multiplex ensembles: Entropy and overlap. Phys. Rev. E. 2013; 87:062806.

12. Bock RD, Aitkin M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika. 1981; 46: 443-459.

13. Bonacich P. Factoring and weighting approaches to status scores and clique identification. Journal of Mathematical Sociology. 1972; 2: 113–120.

14. Borges P. Provas do pré-teste do Enem foram fotografadas no Christus. IG. 2011. Available from: http://ultimosegundo.ig.com.br/educacao/enem/provas-do-preteste-do-enem-foram-fotografadas-no-christus/n1597343813746.html

15. Bradlow ET, Thomas N. Item response theory models applied to data allowing examinee choice. Journal of Educational and Behavioral Statistics. 1998; 23(3): 236–243.

16. Brigham TA. Some effects of choice on academic performance. In Perlmuter, L.C. and Monty, R.A., editors, Choice and perceived control. Hillsdale, NJ: Lawrence Erlbaum. 1979.

17. Cankar G. Allowing Examinee Choice in Educational Testing. Metodološki zvezki. 2010; 2(7):151-166.

18. Cardillo A, Zanin M, Gómez-Gardeñes J, Romance M, García del Amo AJ, Boccaletti S. Modeling the multi-layer nature of the European air transport network: resilience and passengers re-scheduling under random failures. Eur. Phys. J. Special Topics. 2013; 215: 23–33.

19. Casella G, Berger RL. Statistical Inference. 2nd ed. Pacific Grove: Duxbury Press; 2002

20. Chen MH, Shao QM. Monte Carlo estimation of Bayesian credible and HPD intervals. Journal of Computational and Graphical Statistics. 1999; 8(1): 69-92.

21. Chi MTH. Knowledge structures and memory development. In R. S. Siegler (Ed.). Children's thinking: What develops? Hillsdale, NJ: Lawrence Erlbaum Associates, Inc. 1978: 73-96.

22. Chi MTH, Glaser R, Rees E. Expertise in problem solving. In R. J. Sternberg (Ed.). Advances in psychology of human intelligence. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc. 1982: 7-70.

23. Cordova DI, Lepper MR. Intrinsic motivation and the process of learning: beneficial effects of contextualization, personalization, and choice. Journal of Educational Psychology. 1996; 88:715–30.

24. Csardi G, Nepusz T. The igraph software package for complex network research. Inter. Journal Complex Systems. 2006; 1695. Available from: http://igraph.sf.net.

25. Curtis SM. Bugs code for item response theory. Journal of Statistical Software. 2010; 36: 1–34.

26. De Domenico M, Nicosia V, Arenas A, Latora V. Structural reducibility of multilayer networks. Nature Communications. 2015; 6:6864. doi:10.1038/ncomms7864.

27. Dijkstra EW. A note on two problems in connexion with graphs. Numerische Mathematik. 1959; 1:269-271.

28. Donges JF, Schultz HCH, Marwan N, Zou Y, Kurths J. Investigating the topology of interacting networks. Eur. Phys. J. B. 2011; 84:635–651.

29. Fitzpatrick AR, Yen WM. The psychometric characteristics of choice items. Journal of Educational Measurement. 1995; 32:243–259.

30. Fox J, Weisberg S. Nonlinear Regression and Nonlinear Least Squares in R. An appendix to an R companion to applied regression, 2 ed. 2010.

31. Fox JP. Bayesian Item Response Modeling: Theory and Applications. New York, NY: Springer. 2010. ISBN 978-1-4419-0741-7. DOI 10.1007/978-1-4419-0742-4.

32. Freeman LC. Centrality in social networks: conceptual clarification. Social Networks. 1978; 1:215-239.

33. Gilks WR, Thomas A, Spiegelhalter DA. A Language and Program for Complex Bayesian Modelling." The Statistician. 1994; 43: 169-178.

34. Hambleton RK, Swaminathan H. Item response theory: principles and applications. Boston: Kluwer-Nijhoff Publishing; 1985.

35. Hambleton RK, Swaminathan H, Rogers HJ. Fundamentals of Item Response Theory. Newbury Park. CA: Sage; 1991.

36. Hartigan JA, Wong MA. A K-means clustering algorithm. Applied Statistics. 1979; 28: 100–108.

37. Jennings M, Fox J, Graves B, Shohamy E. The test-takers' choice: An investigation of the effect of topic on language-test performance. Language Testing. 1999; 16, 426–456.

38. Kieftenbeld V, Natesan P. Recovery of Graded Response Model Parameters: A Comparison of Marginal Maximum Likelihood and Markov Chain Monte Carlo Estimation. Applied Psychological Measurement. 2012; 36(5): 399-419.

39. Kivelä M, Arenas A, Barthelemy M, Gleeson JP, Moreno Y, Porter MA. Multilayer networks. Journal of Complex Networks. 2014; 2(3): 203-271.

40. Kroese DP, Taimre T, Botev ZI. Handbook of Monte Carlo Methods. Wiley Series in Probability and Statistics. New York: John Wiley and Sons; 2011.

41. Lewis TG. Network Science: Theory and Applications. New Jersey: John Wiley and Sons; 2009.

42. Li W, Liu CC, Zhang T, Li H, Waterman MS, Zhou X J. Integrative analysis of many weighted co-expression networks using tensor computation. PLOS Comput. Biol. 2011; 7:e1001106.

43. Linn RL, Betebenner DW, Wheeler KS. Problem choice by test takers: Implication for comparability and construct validity (CSE Technical Report 485). Los Angeles, CA: University of California, Los Angeles. 1998.

44. Little RJA, Rubin DB. Statistical analysis with missing data. New York, NY: Wiley and Sons; 1987.

45. Lu H-Y. Application of Optimal Designs to Item Calibration. PLoS ONE. 2014; 9(9): e106747. doi:10.1371/journal.pone.0106747.

46. Luce RD, Perry AD. A method of matrix analysis of group structure. Psychometrika. 1949; 14 (2):95–16. doi:10.1007/BF02289146.

47. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS : A Bayesian Modelling Framework: Concepts, Structure, and Extensibility. Statistics and Computing. 2000; 10: 325-337.

48. McDonald JH. Handbook of Biological Statistics. 3rd ed. Baltimore, Maryland: Sparky House Publishing; 2014.

49. Menichetti G, Remondini D, Panzarasa P, Mondragón RJ, Bianconi G. Weighted multiplex networks. PloS one. 2014; 9(6):e97857.

50. Moura RM. Inep amplia pré-teste do Enem e valor do contrato tem aumento de 559%. O Estado de São Paulo. 2010. Available from: http://www.ceaam.net/?sec=71&assunto=& noticia=5410.

51. Patz RJ, Junker BW. Applications and Extensions of MCMC in IRT: Multiple Item Types, Missing Data, and Rated Responses. Journal of Educational and Behavioral Statistics. 1999; 24(4):342-366.

52. Plummer M. JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling.Proceedings of the 3rd International Workshop on Distributed Statistical Computing. Austria, Vienna: In K Hornik, F Leisch, A Zeileis (eds.). 2003. ISSN 1609-395X, URL http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Proceedings/.

53. Powers DE, Bennett RE. Effects of allowing examinees to select questions on a test of divergent thinking. Applied Measurement in Education. 2000; 12: 257–279.

54. Rasch G. Probabilistic models for some intelligence and ttainment test. Copenhagen: Danish Institute for Educational Research; 1960.

55. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria; 2015 . Available from: URL https://www.R-project.org/.

56. Rocklin TR, O'Donnell AM, Holst PM. Effects and underlying mechanisms of self-adapted testing. Journal of Educational Psvchologv. 1995; 87: 103- 116.

57. Rubin, DB. Inference and missing data. Biometrika. 1976; 63(3): 581-592.

58. Siegler RS, Lemaire P. Older and younger adults' strategy choices in multiplication: testing predictions of ASCM using the choice/no-choice method. Journal of Experimental Psychology: General. 1997; 126: 71–92.

59. S.P., Mandelli M. Bando de questões tem menos de 50% do necessário, diz Ministro. O Estado de São Paulo. 2011. Available from: http://www.estadao.com.br/noticias/geral,banco-de-questoes-tem-menos-de-50-do-necessario-diz-ministro-imp-,791524

60. Spiegelhalter D, Thomas A, Best N, Lunn D. WinBUGS User Manual Version 1.4. MRC Biostatistics Unit. 2003. URL http://www.mrc-bsu.cam.ac.uk/wp-content/uploads/manual14.pdf.

61. Thomas A, O'Hara B, Ligges U, Sturtz S. Making BUGS Open. R News, 2006; 6(1): 12-17. URL http://CRAN.R-project.org/doc/Rnews/.

62. Van der Linden WJ, Ren H. Optimal Bayesian Adaptive Design for Test-Item Calibration. Psychometrika. 2014; 1–26.

63. Verbrugge LM. Multiplexity in adult friendships. Soc. Forces. 1979; 57:1286–1309.

64. Wainer H, Wang XB, Thissen D. How well can we compare scores on test forms that are constructed by examinees' choice? Journal of Educational Measurement. 1994; 31:183 – 199.

65. Wang WC, Jin KY, Qiu XL, Wang L. Item Response Models for Examinee-Selected Items. Journal of Educational Measurement. 2012; 49(4):419–445.

66. Wang X, Wainer H, Thissen D. On the viability of some untestable assumptions in equating exams that allow examinee choice. Applied Measurement in Education. 1995; 8:211-225.