Multistage adaptive testing based on logistic positive exponent model

Thales Akira Matsumoto Ricarte

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura:_____
_

Thales Akira Matsumoto Ricarte

Multistage adaptive testing based on logistic positive exponent model

Doctoral dissertation submitted to the Instituto de Ciências Matemáticas e de Computação - ICMC-USP and to the Departamento de Estatística – DEs-UFSCar, in partial fulfillment of the requirements for the degree of the Doctorate Joint Graduate Program in Statistics DEs-UFSCar/ICMC-USP. *EXAMINATION BOARD PRESENTATION COPY*

Concentration Area: Statistics

Advisor:Profa. Dra. Mariana Cúri

Coadvisor: Prof. Dr. Jorge Luis Bazán Guzmán

USP – São Carlos
November 2016

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

A487m Akira Matsumoto Ricarte, Thales
Multistage Adaptive Testing based on Logistic Positive Exponent model / Thales Akira Matsumoto Ricarte; orientadora Mariana Cúri; co-orientador Jorge Luis Bazán Guzmán. -- São Carlos, 2016.
70 p.

Tese (Doutorado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2016.

1. Logistic Positive Exponent. 2. Multistage Adaptive Testing. 3. Fisher Information. 4. Kullback-Leibler Information. 5. Continuous Entropy Method. I. Cúri, Mariana , orient. II. Bazán Guzmán, Jorge Luis, co-orient. III. Título.

Thales Akira Matsumoto Ricarte

Teste adaptativo multiestágio baseado no modelo logístico de expoente positivo

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Doutor em Estatística – Interinstitucional de Pós-Graduação em Estatística. *EXEMPLAR DE DEFESA*

Área de Concentração: Estatística

Orientadora: Profa. Dra. Mariana Cúri

Coorientador: Prof. Dr. Jorge Luis Bazán Guzmán

USP – São Carlos
Novembro de 2016

I dedicate this thesis to my family and friends.

ACKNOWLEDGEMENTS

I want to thank you the following people for the execution of this thesis.

To my advisors, Mariana Cúri, Alina von Davier and Jorge Bazán. Not only for your advices and support, but also for the opportunity to experience another country.

To my parents, Satie and Juarez, to my siblings, Rosana and Fabio and my dogs Piti (that, unfortunately, died in 2015), Mordida and Nanica, for accompanying during the years.

To Noemi, Cabeća, Jackson, Juan, Amélia and Katherine, for all support and fun times.

To Jonas and Masha, for the 9 months I stayed in their house during my Sandwich program.

To Jessica, Claudia, Jiangang, Yoav, Mengxiao, Saad, Yanchi, Rubi, Duanli and all other ETS members that I met during my wonderful stay in Princeton, NJ.

To ICMC institute for the structure and support given to me over the years.

To CNPQ and CAPES for the financial support.

ABSTRACT

RICARTE, T. A. M.. **Multistage adaptive testing based on logistic positive exponent**. 2016. 70 f. Doctoral dissertation (Doctorate Candidate em joint Graduate Program in Statistics DEs-UFSCar/ICMC-USP) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

The Logistic Positive Exponent (LPE) model from Item Response Theory (IRT) and the Multistage Adaptive Testing (MST) using this model are the focus of this dissertation. For the LPE, item parameter estimations efficiency was studied, it was also analyzed the latent trait estimation for different response patterns to verify the effects it has on guessing and accidental mistakes. The LPE was put in contrast to Rasch, 2 and 3 parameter logistic models to compare the its efficiency. The item parameter estimations were implemented using the Bayesian approach for the Monte Carlo Markov Chain and the Marginal Maximum Likelihood. The latent trait estimation were calculated by the Expected a Posterior method. A goodness of fit analysis were made using the Posterior Predictive model-check method and information statistics. In the MST perspective, the LPE was compared with the Rasch and 2 logistic models. Different tests were constructed using methods that uses optimization functions to select items from a bank. Three functions were chosen to this task: the Fisher and Kullback-Leibler informations and the Continuous Entropy Method. The results were obtained with simulated and real data, the latter was from a general science knowledge test calls General Science test and it was provided by the Educational Testing Service company. Results showed that the LPE might help individuals that made mistakes in earlier stage of the test, especially for easy items. However, the LPE requires a large individual sample and time to estimate the item parameters making it an expensive model. MST based on LPE can be dissolve the impact of accidental mistakes from high performance test takers depending of the item pool available and the way the test is constructed. The optimization function performance vary depending of the situation.

Key-words: Logistic Positive Exponent, Multistage Adaptive Testing, Fisher Information, Kullback-Leibler Information, Continuous Entropy Method.

RESUMO

RICARTE, T. A. M.. **Multistage adaptive testing based on logistic positive exponent**. 2016. 70 f. Doctoral dissertation (Doctorate Candidate em joint Graduate Program in Statistics DEs-UFSCar/ICMC-USP) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

O modelo Logístico de Expoente Positivo (LPE) da Teoria de Resposta ao Item (IRT) e o Teste Adaptativo Multiestágio (MST) sob esse modelo são os focos desta tese. Para o LPE, a eficiência da estimaçôes dos parâmetros dos itens foram estudados, também foi analisado como as estimativas dos parâmetros dos indivíduos foram influênciados por padrões de respostas contendo chutes ou erros acidentais. O LPE foi comparado com os modelos de Rasch, Logístico de 2 e 3 Parametros para verificar seu desempenho. A estimação dos pârametros dos itens foi implementada usando Monte Carlo via cadeias de Markov sob a abordagem Bayesiana e a Maxíma Verossimilhança Marginal. As estimações dos traços latentes foram calculadas atravéz do Método da Esperança a Posteriori. A qualidade do ajuste dos modelos foram analisadas usando o método *Posterior Predictive model-check* e critério de informações. Sob o contexto do MST, o LPE foi comparado com os modelos de Rasch e Logístico de 2 Parametro. Os MSTs foram contruidos usando diferentes funções de objetivas que selecionaram os itens de bancos para comporem os testes. Três funções foram escolhidas para esse trabalho: As informações de Fisher e Kullback-Leibler e o *Continuous Entropy Method*. Os resultados para dados simulados e reais foram obtidos, os dados reais eram consituidos de respostas a perguntas sob conhecimento científico de do *General Science test* que foram fornecidos pela empresa *Educational Testing Service*. Resultados mostraram que o LPE pode ajudar os indivíduos que cometeram erros acidentais nas primeiras perguntas do teste, especialmente para os itens fáceis. Entretanto, este modelo requer tempo e uma grande quantidade de amostras de indivíduos para calcular as estimativas dos parâmetros dos itens o que o torna um modelo caro. O MST sob o modelo LPE pode diminuir o impacto de erros acidentais cometidos por examinandos com alto desempenho dependendo dos itens disponivéis no banco e a forma de construção do MST. O desempenho das funções objetivas varianram de acordo com cada situação.

Palavras-chave: Logístico de Expoente Positivo, Teste Adaptativo Multiestágio, Informação de Fisher, Informação de Kullback-Leibler, Continuous Entropy Method.

LIST OF FIGURES

LIST OF TABLES

CONTENTS

CHAPTER

1

INTRODUCTION

Assessments are important tools to measure educational performance. Test results can assist students, teachers and even nations to implement course of actions to improve their education. Exams also helps with selections of individuals for college, scholarship programs and jobs. However, valid tests can be very time consuming, causing fatigue to the examinees, and costly.

It is clear the importance of having tests that are reliable and efficient. In this perspective, the psychometric researchers are constantly studying and developing methodologies to enhance the test measurement efficiency (better latent trait estimates, shorter tests, choosing adequate items for the test, etc.). Two of those techniques are Item Response Theory (HAMBLETON; SWAMINATHAN; ROGERS, 1991, IRT), which is used for constructing tests and analyzes the relationship of individual and items characteristics, and Multistage Adaptive Testing (YAN; DAVIER; LEWIS, 2014, MST), a method for test optimization.

The IRT models relate the individuals' latent traits with the probability of selecting an item response category. Latent traits are attributes that are not directly measured, for example: knowledge on a subject or creativity. The curve that describes the probability of choosing a response category of an item and the latent trait is called Item Characteristic Curve (ICC). Graphics of ICCs are shown in the next chapter in Figures 1 and 2.

A wide variety of models were developed for different items and latent trait characteristics. The Rasch, 2 and 3 Parameters Logistic models (2PL and 3PL, respectively) are the most popular ones. These models are suitable to describe dichotomous items and continuous unidimensional latent trait.

Due to well known advantages of IRT models Vianna (1982), Linden e hambleton (1997), they are already applied in some high stake tests. Examples of those tests are: The Test of English as a Foreign Language, TOEFL ETS (2015b), an language proficiency test applied worldwide; The Graduate Record Examination, GRE ETS (2015a), which is required for many graduate

schools in USA; and the Brazilian's National High School Exam, ENEM INEP (2016), used as criteria in many scholarship programs and college admission process in Brazil.

Since high stake tests are of great relevance, they also can induce a high level of stress in candidates, resulting in possible mistakes in easier questions, especially in earlier stages of the test Barton e Lord (1981). Depending on the IRT model, these errors can have different effects on the latent trait estimates.

In 3PL model, it could be difficult to an individual to recover from mistakes to easier items. In 2PL model, latent trait estimates are not affected by the difficulty parameter values, meaning that mistakes on items with different difficulty parameters will have the same impact on the latent trait estimates if they have the same discrimination parameter values. In Rasch model, the proportion of correct responses is a sufficient statistic for latent trait.

Samejima (2000) proposed the Logistic Positive Exponent (LPE) family of models that generalizes the 2PL model by adding an exponential parameter, providing a better methodology for understanding the complexity of human behavior. Depending on the exponential parameter value, the model can penalize wrong response to easy items (similar to 3 Parameter Logistic model) or it can facilitate the recovery from earlier mistakes in the test (can be good for stress situations).

The asymmetry ICC of LPE model is resultant of an exponential factor added to the 2PL. Some others IRT models with asymmetric ICC with different approaches can be found in: Bolfarine e Bazan (2010a) who used scobit Achen (2002) as the link function instead the logistic's, Bazán *et al.* (2004) who build the asymmetric ICC using the skew-normal distribution Azzalini (2005), Molenaar, Dolan e Boeck (2012) and Molenaar (2015) who proposed heteroscedastic latent trait models.

There are already previous research using the LPE model in the literature made by other authors.

In Bolfarine e Bazan (2010a), the model was applied to 974 fourth-grade Peruvian students responses to a 18 items Math test. In their results, the LPE had a better fit to the data than the symmetric models.

Santos, Gamerman e Soares (2012) added the guessing parameter to the LPE. In her dissertation, she used both the skew-normal and skew-logistic to detect asymmetric items, and it was observed that the asymmetric items had convergence troubles.

Another modification was presented by Flores (2012) by adding a testlet effect in the LPE model to address local item dependency problems. The model was applied to responses of 2nd grade elementary students in Peru to the Reading Comprehension and Mathematics of the Census Student Testing from 2010. Her conclusions were that models without testlet effect overestimated the latent traits.

Bolt, Deng e Lee (2014) utilized the LPE model in a vertical scaling context, where the objective is to analyze the progress of students during a period of time, that can also be the case where the same test is applied to individuals on different grades. They shown that the LPE-related misspecification are difficult to detect when utilizing the 2PL and 3PL model, and this misfit can masquerade the growth estimation. They also used real data from Wisconsin Knowledge & Concepts Examination Mathematics tests from 2 years to compare the LPE (they also added a guessing parameter), 2PL and 3PL model fit. A random sample of 1000 individual responses to the 46 items were used for each test. The LPE usually had lower Deviance information criterion (DIC) than the 2PL and 3PL, indicating that LPE had better fit.

In these studies, the exponential factor introduced in the LPE model caused several practical issues such as problems of identifiability and high correlation between parameters, resulting in a necessity of high sample of individuals, time expensive, difficulty in convergence, etc.

It is on the interest of this dissertation to study the LPE model, compare with Rasch, 2PL and 3PL models and evaluate the fitness of the model on a real data set. For the latter, it was implemented the Posterior Predictive Model-Check (PPMC), a method based on comparing frequencies of real and generated data, and information statistics: Expected Akaike Information Criterion (BROOKS, 2002, E-AIC), Expected Bayesian Information Criterion (E-BIC), and DIC (E-AIC, E-BIC, and DIC Spiegelhalter *et al.* (2002).

To secure a certain degree of reliability and efficiency, linear tests could be very long, since there are no optimization involved. One alternative can be adaptive tests, where items are present to the individual according to their previous responses. An example could be the Computarized Adaptive Testing (LINDEN; GLAS, 2000, CAT) that applies personalized tests by administering items to individuals according the their previous responses. This process is made through a item selection procedure that choose one item after each response given. The customization causes CAT to have reduced length by losing some estimation efficiency. Both TOEFL and GRE tests mentioned above have or had CAT versions.

However, CAT may have practical issues and be difficult to implement. Since CAT administers item-by-item, item review is usually not a feasible and it is hard to validate because of the high number of test possibilities. Moreover, the item selection and exposure control can be very complex.

Those issues can be minimized by implementing the MST instead. The GRE is currently using this alternative.

An MST is composed by pre-assembled subsets of items called modules. These modules are presented to the examinee in an adaptive manner by selecting new modules according to the individual responses to previous ones. This process is not as good as CAT's in terms of estimation efficiency and test length, but MST is easier to implement than a CAT. In test application, IRT

models can be implemented to support both CAT and MST structure.

Those modules can be assembled by maximizing/minimizing an objective function using linear programming.

In CAT, the Fisher and Kullback-Leibler (KL) information are commonly used as objective function in the item selection criteria. This concept can be extended for the MST's module assembly as well. Similarly, the Continuous Entropy method (CEM), which was implemented in cognitive diagnostic CAT Cheng (2009), can also be used.

In the literature, it is common to find the module assembly methods based on Fisher information. For the Rasch and 2PL model, items have the most information when difficulty parameter and latent trait have the same value. This situation corresponds to 50% of probability of correct response. However, 50% probability of correct response does not implies the equality of difficulty and ability for the LPE model. This fact contradicts Lord (1970), who defends that a test should administer items that are neither too difficult nor too easy for examinees , i.e., individuals would have approximately 50% probability to give the correct response for the item in the cases of dichotomous IRT models. Thus, it is on the interest of this study to analyze different optimization functions.

1.1 Objective and Organization

The objective of this dissertation are:

- To analyze the advantages and disadvantages of a particular case of LPE in terms of parameter estimation and model fit
- To show the similarity and differences of this asymmetric model and other logistic models (Rasch, 2PL and 3PL)
- To verify its performance on simulated and a real datasets
- To implement and analyze the performance of MST's under the LPE model
- To compare the LPE model in relation to Rasch and 2PL based MSTs
- To analyze the effects of LPE based MST when guessing and mistakes are made.
- To evaluate which optimization function (Fisher and KL informations and CEM) have better results, in terms of latent trait estimation and probability of correct response, for the LPE based MST.

The dissertation is organized in the following manner: chapter 2 introduces the LPE model, its Item Characteristic Curve (ICC) and a particular case, due to estimation issues, of the model was

adopted. In chapter 3, a frequentist and a Bayesian estimation methods for item parameters and a Bayesian approach for the latent trait are presented. In chapter 4, the Multistage Adaptive test is introduced alongside with the Fisher and KL informations and CEM in the context of module assembly methods. Additionally, a comparison among these functions under the LPE model using are made. In chapter 5, the Posterior Predictive Model-Check and the information statistics used in this article, to analyze the model fit, are introduced. In chapter 6, three results using the Rasch, 2PL and LPE models are shown in three sections: section 6.1, shows simulations about how the LPE asymmetry results in different response patterns, section 6.2 presents a comparison of the frequentist and Bayesian performance in the item parameter estimations and section 6.3 shows the application of the model using a real dataset. In chapter 7, results of the implementation of MSTs under Rasch, 2PL and LPE models using simulated (sections 7.1, 7.2 and 7.3) and real data (section 7.4) and the methodology in previous sections are made. Furthermore, the performance of Fisher and KL informations and CEM are analyzed. Finally, in chapter 8, discussions, conclusions and future studies of the methodology adopted in this dissertation are presented.

CHAPTER

2

THE LPE MODEL

Samejima (2000) pointed out that some items requires multiple steps to be solved and if theses steps are hard to execute, they can increase the difficulty of an item unevenly. The probability of correct response for individuals with high latent trait can be lower in these cases than on a single process item. However, individuals with low latent traits would be much more penalized since they are would have to complete all the steps to solve these items. Samejima showed that in these scenarios, it might be inappropriate to assuming that these items have symmetric ICCs.

$$P(X_{ij}=1 \mid \theta_i, a_j, b_j, \lambda_j) = P(X_{ij} \mid \theta_i, \boldsymbol{\zeta}_j) = \left[\frac{1}{1+\exp(-a_j(\theta_i - b_j))}\right]^{\lambda_j}, \tag{2.1}$$

where X_{ij} is a binary random variable that assumes value of 1 if the examinee with latent trait $\theta_i, i \in \{1,...,I\}$, choses the correct response for the item $j \in \{1,...,J\}$, and 0 otherwise; $P(X_{ij}=1 \mid a_j, b_j, \lambda_j, \theta_i)$ is the probability of the examinee to correctly respond the item; $a_j > 0$, b_j, $\lambda_j > 0$ are the discrimination, the difficulty and the acceleration parameters, respectively, and $\boldsymbol{\zeta}_j = (a_j, b_j, \lambda_j)$.

The 2PL model, $P_{2pl}(X_{ij}=1 \mid \theta_i, a_j, b_j)$, can be obtained by fixing the $\lambda = 1$ in (2.1), and if a parameter is also fixed to 1, the Rasch model is obtained. Analogously to LPE, the 3PL model is also a generalization of the 2PL model: $c + (1-c)P_{2pl}(X_{ij}=1 \mid \theta_i, a_j, b_j)$, where $0 \leq c \leq 1$ is the guessing parameter.

In Samejima (1997) and in chapter 6 of this dissertation, it is shown that, contrary to the 2PL, for the asymmetric items under the LPE ($\lambda \neq 1$), the difficulty parameter affects the θ estimate. Wrong answers to easier items with $\lambda < 1$ have greater negative impact in the individual's ability estimate as the 3PL, while for items with $\lambda > 1$, right answers to more difficult items have greater positive impact on the individual's ability estimate, which is not contemplated in the 3PL.

Another interesting view of the LPE model is that, for low λ values, the ICC can have a pseudo-lower asymptote, resulting in a similarity to the 3PL model. This effect can be visualized in the Figure 1. For this example, the LPE item parameters were fixed at: $a = 1$, $b = 2$ and $\lambda = 0.18$, and the 3PL parameters were fixed at: $a = 0.5$, $b = -2$ and $c = 0.1$.

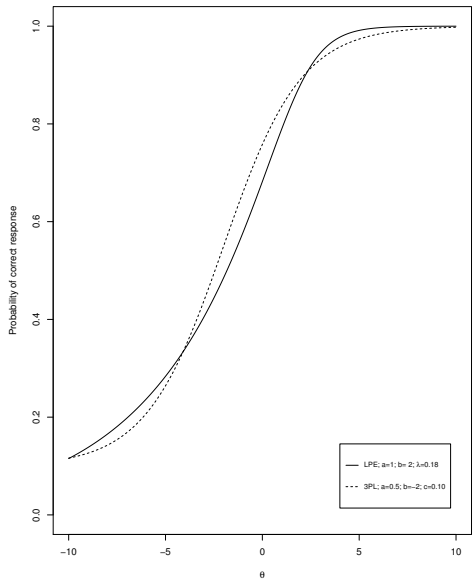


Figure 1 – LPE ICC with 3PL similarity

Note in Figure 1 that, even though the two curves are similar, the parameter values are very different, especially the b. In those two models, the probability of correct response = 0.5 does not occours at $\theta_{50\%} = b$, but at $\theta_{50\%} = b + \frac{ln(1-2c)}{a}$ and $\theta_{50\%} = b - \frac{ln(\sqrt[\lambda]{2}-1)}{a}$ for 3PL and LPE, respectively.

However, there is a downside to this model: the estimation of the item parameters concomitantly is not straight forward, requiring a large sample size and time. The reason is that both a and λ parameters influences on the ICC's inclination, and both b and λ modify the positioning of the curve leading to a problem of identifiability and difficulty in estimation. For this reason, in this dissertation, the a parameter is fixed at 1 for all items to facilitate the estimation process. We will refer the model with this constraint as the “Rasch" LPE model.

In Figure 2, it is shown three Item Characteristic Curves (ICC), 2 for the “Rasch" LPE and one for the Rasch model, b parameter are equal to 0.

For $\lambda = 0.5$, the ”Rasch" LPE ICC is on the right from Rasch's and the curve is more steep. For $\lambda = 2$, the ”Rasch" LPE ICC is on the left from Rasch's and the curve is less steep.

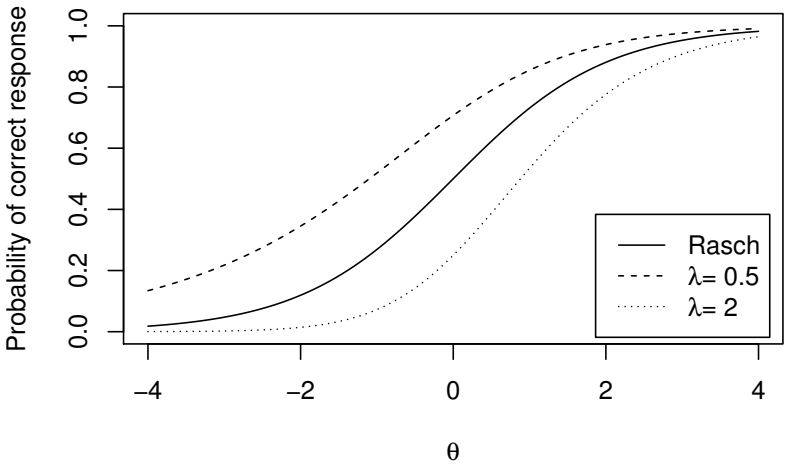


Figure 2 – Examples of "Rasch" LPE ICC's with *b* parameters = 0

CHAPTER

3

ESTIMATION METHODS

When implementing a test using IRT models, it is common to have two estimation stages. The first stage, usually called item calibration, consists in estimating the item parameters. In this stage, the individual's parameters may or may not be known. In this article, the latent traits in this stage is considered unknown. The second stage consists on the estimation of individual's latent traits.

For the first stage, there are several approaches that can be used to obtain the item parameters' when the θ is not known. In our case, 2 methods were used: Markov Chain Monte Carlo (MCMC) in a Bayesian context and the Marginal Maximum Likelihood (MML) Bock e Aitkin (1981), Thissen (1982), Rigdon e Tsutakawa (1983), Bock, Gibbons e Muraki (1988), Wilson, Wood e Gibbons (1991). The MCMC and MML are described in section 3.1.

The Bayesian MCMC does not have problems with aberrant responses and it is easier to implement. In the MML method, the item parameters are estimated without the latent trait, thus avoiding problems that would surge in the Joint Maximum Likelihood estimation. In chapter 6, comparisons between both methods for the "Rasch" LPE model are presented.

The second stage is focused on the latent trait estimation with item parameters considered known or fixed to the estimates of stage one. Methods used in this stage include maximum likelihood, Mode a Posteriori and Expected a Posteriori. For our results, the method used to do these estimation is the Expected a Posteriori (EAP) presented in subsection 3.2. The EAP was chosen because it does not require the implementation of an iterative algorithm to find the estimates because just a simple numeric integration is needed.

3.1 Item Calibration

For item calibration, a pre-tests is administered to a sample of the target population's, The latter is assumed to have a certain distribution. The MCMC and MML are described in

subsections 3.1.1 and 3.1.2, respectively.

3.1.1 Bayesian approach and MCMC method

Bayesian modeling state that the parameters distribution is a combination of data and a prior knowledge of the area. According to the Bayes theorem the posterior distribution of the individuals' parameters $\boldsymbol{\theta}$ and item parameters $\boldsymbol{\zeta}$ are written as

$$\pi(\boldsymbol{\theta},\boldsymbol{\zeta} \mid \boldsymbol{X}) \propto L(\boldsymbol{X} \mid \boldsymbol{\theta},\boldsymbol{\zeta}) f(\boldsymbol{\theta},\boldsymbol{\zeta}), \tag{3.1}$$

where $L(\boldsymbol{X} \mid \boldsymbol{\theta})$ is the likelihood function (data) and $f(\boldsymbol{\theta},\boldsymbol{\zeta})$ is called the prior distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ (prior knowledge).

Considering the "Rasch" LPE model under the Bayesian approach, let $\mathbf{X} = [X_{ij}]$, a binary random variable matrix, and X_{ij}, i and j as defined in (2.1). Because in IRT it is assumed the independence among individual responses and assuming local item independence, the likelihood function is given by

$$L(\boldsymbol{X} \mid \boldsymbol{\theta},\boldsymbol{\zeta}_{lpe}) = \prod_{i=1}^{I}\prod_{j=1}^{J} P(X_{ij} \mid \theta_i,\boldsymbol{\zeta}_j)^{X_{ij}} Q(X_{ij} \mid \theta_i,\boldsymbol{\zeta}_j)^{(1-X_{ij})}, \tag{3.2}$$

where $P(X_{ij} \mid \theta_i,\boldsymbol{\zeta}_j)$ is the "Rasch" LPE model presented in the (2.1) with $\lambda = 1$, $Q(X_{ij} \mid \theta_i,\boldsymbol{\zeta}_j) = 1 - P(X_{ij} \mid \theta_i,\boldsymbol{\zeta}_j)$, $\boldsymbol{\theta} = \{\theta_1,...,\theta_I\}$ is the vector of latent traits with θ_i the latent trait of individual i and $\boldsymbol{\zeta}_{lpe} = \{\boldsymbol{\zeta}_1,...,\boldsymbol{\zeta}_J\}$ is the set of its item parameters with $\boldsymbol{\zeta}_j = \{b_j,\lambda_j\}$.

Then, the LPE posteriori distribution is written as

$$\pi(\boldsymbol{\theta},\boldsymbol{\zeta}_{lpe} \mid \mathbf{X}) = \frac{L(\boldsymbol{X} \mid \boldsymbol{\theta},\boldsymbol{\zeta}_{lpe}) f(\boldsymbol{\theta},\boldsymbol{\zeta}_{lpe})}{\int\int L(\boldsymbol{X} \mid \boldsymbol{\theta},\boldsymbol{\zeta}_{lpe}) f(\boldsymbol{\theta},\boldsymbol{\zeta}_{lpe}) d\boldsymbol{\theta} d\boldsymbol{\zeta}_{lpe}}, \tag{3.3}$$

where $f(\boldsymbol{\theta},\boldsymbol{\zeta}_{lpe})$ is the joint prior distribution for individual and item parameters. In this model, the assumption of local independence between items and between $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}_{lpe}$ are made, meaning that $f(\boldsymbol{\theta},\boldsymbol{\zeta}_{lpe}) = f(\boldsymbol{\theta}) f(\boldsymbol{\zeta}_{lpe})$, where $f(\boldsymbol{\theta}) = \prod_{i=1}^{I} f(\theta_i)$ and $f(\boldsymbol{\zeta}_{lpe}) = \prod_{j=1}^{J} f(b_j) f(\lambda_j)$.

One way to estimate the parameters is to apply the MCMC method. The idea of this method is to create a Markov Chain M_0, M_1, M_2,...,M_K, which represents a state $M_k = (\boldsymbol{\theta}^k,\boldsymbol{\zeta}^k); k \in \{1,...,K\}$, $\boldsymbol{\theta}^k$ and $\boldsymbol{\zeta}^k$ are the latent trait and item parameter of the chain kth iteration, respectively.

To build the chain, the Metropolis-Hasting algorithm can be implemented. In this iterative algorithm, at each iteration a transition distribution is used to generate candidates for the chain and through an acceptance or rejection rule, the k-th chain values of the iteration are going to assume values from the candidates or from the previous iteration. This procedure is done until

the chain reaches the desired target distribution (in our case, the posteriori distribution). Then, the chain can be used to infer about the model parameters.

A pseudo-algorithm of the Metropolis-Hastings is shown below.

1) starting values $M_0 = (\boldsymbol{\theta}^0, \boldsymbol{\zeta}^0)$, in which $\pi(\boldsymbol{\theta}^0, \boldsymbol{\zeta}^0 \mid \mathbf{X}) > 0$, are assigned for each unknown parameter in the model.

2) next, a candidate, $M^* = (\boldsymbol{\theta}^*, \boldsymbol{\zeta}^*)$, is sampled from a transition distribution $T_k(\boldsymbol{\zeta}^* \mid \boldsymbol{\zeta}^{k-1})$.

3) Then, attribute $M^k = M^*$ with probability $r = \frac{\pi(\boldsymbol{\zeta}^* \mid \mathbf{X}) T_k(\boldsymbol{\zeta}^{k-1} \mid \boldsymbol{\zeta}^*)}{\pi(\boldsymbol{\zeta}^{k-1} \mid \mathbf{X}) T_k(\boldsymbol{\zeta}^* \mid \boldsymbol{\zeta}^{k-1})}$, or else, maintain $M^k = M^{k-1}$.

4) repeat 2) and 3) until the convergence is reached.

Under certain regularity conditions Tierney (1994), the chain will converge to the target $\pi(\boldsymbol{\theta}, \boldsymbol{\zeta})$ for large k.

3.1.2 The Marginal Maximum Likelihood (MML)

Bock e Lieberman (1970) proposed a Marginal Maximum Likelihood approach to estimate the item parameters on a Normal-Ogive IRT model. In their work, the marginalized probability of a response vector is reached by integrating the distribution of each individual response vector over the θ distribution. If this distribution is correctly specified, the estimates of the item parameter are going to be consistent.

Let $X_{i.} = \{X_{i1}, \ldots, X_{iJ}\}$ be the response vector of the individual i for the J items in the test. The probability of a response vector for an individual i is given by

$$P(X_{i.} \mid \theta_i, \boldsymbol{\zeta}_{lpe}) = \prod_{j=1}^{J} P(X_{ij} \mid \theta_i, \boldsymbol{\zeta}_j)^{X_{ij}} Q(X_{ij} \mid \theta_i, \boldsymbol{\zeta}_j)^{(1-X_{ij})}, \tag{3.4}$$

where $P(X_{ij} \mid \theta_i, \boldsymbol{\zeta}_j)$, $Q(X_{ij} \mid \theta_i, \boldsymbol{\zeta}_j)$, X_{ij}, θ_i, $\boldsymbol{\zeta}_{lpe}$ and $\boldsymbol{\zeta}_j$ are defined in (3.2).

The marginalized probability of the response vector over θ_i is defined as:

$$P(X_{i.} \mid \boldsymbol{\zeta}_{lpe}) = \int P(X_{i.} \mid \theta_i, \boldsymbol{\zeta}_{lpe}) f(\theta_i) d\theta_i, \tag{3.5}$$

where $f(\theta_i)$ is the prior distribution for θ_i. In IRT it is usual to consider θ_i having a $Normal(\mu, \sigma^2)$ distribution, with μ and σ^2 as the mean and variance, respectively. For this article, $\mu = 0$ and $\sigma^2 = 1$

and the marginal likelihood function is written as

$$L(\boldsymbol{X} \mid \boldsymbol{\zeta}_{lpe}) = \prod_{i=1}^{I} P(X_{i.} \mid \boldsymbol{\zeta}_{lpe}).$$

and its the logarithm is given by

$$l(\boldsymbol{X} \mid \boldsymbol{\zeta}_{lpe}) = \sum_{i=1}^{I} \log(P(X_{i.} \mid \boldsymbol{\zeta}_{lpe})). \tag{3.6}$$

The MML estimator for the LPE model is obtained through solving the equations originated from the first partial derivatives of (3.6) in relation of each item parameter of the LPE model equalized to 0.

The first derivate of the equation (3.6) in relation to each item parameter is given by (More details in Bock and Lieberman, 1970)

$$\frac{\partial}{\varsigma_j} l(\boldsymbol{X} \mid \boldsymbol{\zeta}_j) = \sum_{j=1}^{J} \pi(\theta_i, \boldsymbol{\zeta}_j \mid \mathbf{X}) \prod_{h \neq j}^{J} [P(X_{ih} \mid \theta_i, \boldsymbol{\zeta}_h)^{X_{ih}} Q(X_{ih} \mid \theta_i, \boldsymbol{\zeta}_h)^{(1-X_{ih})}](-1)^{X_{ih}+1} \frac{\partial}{\varsigma_j} P(X_{ij} \mid \theta_i, \boldsymbol{\zeta}_j). \tag{3.7}$$

where $\pi(\theta_i, \boldsymbol{\zeta}_j)$ is defined in (3.3) for $\boldsymbol{\theta} = \theta_i$ and $\varsigma_j \in \{b_j, \lambda_j\}$. The derivate $\frac{\partial}{b_j} P(X_{ij} \mid \theta_i, \boldsymbol{\zeta}_j) = -\lambda_j(1+\exp(-(\theta_i + -b_j)))^{-\lambda_j - 1} \exp(-(\theta_i - b_j))$ and $\frac{\partial}{\lambda_j} P(X_{ij} \mid \theta_i, \boldsymbol{\zeta}_j) = \log(1+\exp(-(\theta_i - b_j))(1+\exp(-(\theta_i - b_j))^{\lambda_j}$.

Finding the MML estimators from (3.7) involves solving integrals that don't have a known analytical solution. One way to solve this problem is to use quadratures to numerically approximates its value. Let Y_1,...,Y_K be the midpoint of K rectangles that subdivides a region of θ's parametric space. Using quadratures method in (3.7), the MML estimates for each parameter can be calculate by finding the solutions of the following equations:

$$\hat{b}_j \equiv \sum_{k=1}^{K} \left(\frac{\bar{r}_{jk} - P_j(Y_k)\bar{\eta}_{jk}}{P_j(Y_k)Q_j(Y_k)} \right) \lambda_j (1 + \exp(-(Y_k - b_j)))^{-\lambda - 1} \exp(-(Y_k - b_j)) = 0,$$

$$\hat{\lambda}_j \equiv \sum_{k=1}^{K} \left(\frac{\bar{r}_{jk} P_j(Y_k)\bar{\eta}_{jk}}{Q_j(Y_k)} \right) (-\log(1 + \exp(-(Y_k - b_j)))) = 0,$$

where $P_j(Y_k) = P(X_{ij} \mid \theta_i = Y_k, \zeta_j)$, $Q_j(Y_k) = 1 - P_j(Y_k)$, $\bar{r}_{jk} = \sum_{i=1}^{n} \frac{\prod_j^J X_{ij} P_j(Y_k)^{X_{ij}} Q_j(Y_k)^{1-X_{ij}} A(Y_k)}{\sum_{k=1}^{K} \prod_j^J X_{ij} P_j(Y_k)^{X_{ij}} Q_j(Y_k)^{1-X_{ij}} A(Y_k)}$, $\bar{\eta}_{jk} = \sum_{i=1}^{n} \frac{\prod_j^J P_j(Y_k)^{X_{ij}} Q_j(Y_k)^{1-X_{ij}} A(Y_k)}{\sum_{k=1}^{K} \prod_j^J X_{ij} P_j(Y_k)^{X_{ij}} Q_j(Y_k)^{1-X_{ij}} A(Y_k)}$ and $A(Y_k)$ is the probability of the standard Normal distribution at Y_k.

3.2 Latent trait parameters

The Expected a Posteriori estimation method consists in calculation of the expected value of he parameters' posterior distribution. The EAP method to estimate the latent under an unidimensional IRT model is described below.

Let i be the index of the ith individual with latent trait θ_i and $j \in \{1,...,J\}$ be the index of the items, the EAP estimation method is given by:

$$\hat{\theta}_i = \frac{\int \theta_i L(X \mid \theta_i, \boldsymbol{\zeta}_{lpe}) f(\theta_i) d\theta_i}{\int L(X \mid \theta_i, \boldsymbol{\zeta}_{lpe}) f(\theta_i) d\theta_i}, \tag{3.8}$$

where, $L(X \mid \theta_i, \boldsymbol{\zeta}_{lpe}) = \prod_{j=1}^{J} P(X_{ij} \mid \theta_i, \boldsymbol{\zeta}_j)^{x_{ij}} * Q(X_{ij} \mid \theta_i, \boldsymbol{\zeta}_j)^{1-x_{ij}}$ is the likelihood function, $P(X_{ij} \mid \theta_i, \boldsymbol{\zeta}_j)$ is an IRT model, $Q(X_{ij} \mid \theta_i, \boldsymbol{\zeta}_j) = 1 - P(X_{ij} \mid \theta_i, \boldsymbol{\zeta}_j)$, $\boldsymbol{\zeta}_j$ is a vector of known item parameters for the item j, $f(\theta_i)$ and X_{ij} as defined in (3.2) and (3.5), respectively.

As a Bayesian method, the prior standard normal distribution for parameter θ's is assumed, largely employed in educational measurement and psychometric fields (need reference), thus avoiding problems caused by extreme response patterns.

CHAPTER

4

MULTISTAGE TEST

A Multistage Adaptive Test (MST) is a test composed by pre-assembled short linear tests called *modules* and is administered in stages (minimum of 2). These modules have different levels of difficulty. The adaptive part of the test comes due the fact that, at each stage, a module is selected to the individual according to their performance on previous stages (example of MST in Figure 3). In MST, the modules in the first stage (usually, there is only one module in stage 1) are called Routing modules and the module selection criteria is called Routing.

The modularized structure of this type of test allows for an easier way to validate the test beforehand and it also facilitate the implementation of item revision within each module Yan, Davier e Lewis (2014).

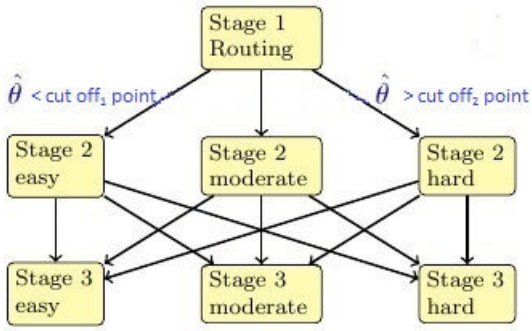
The structure compounded by the modules, stages and the Routing is called *panel*. A multistage test can be formed by multiple parallel panels to improve security and exposure rate. When an individual is going to take a test, one of the panels is selected, and the test taker respond a subset of the modules defined by a path within this panel. In this dissertation, only one panel was built for each MST.

The number of panels, stages, modules per stage and items per module on an MST depends on the the test's purpose. Veldkamp (2014) describes a method to design a blueprint for the MST, this facilitates the elaboration of an item bank (instead of using an existing one) following an MST structure by having information about the number of items with certain characteristics needed. This dissertation does not use the design of a blueprint, but this method can be very useful while implementing a multistage test in a real application.

Figure 3 shows a diagram of a three stage panel with a Routing module on stage 1, and 3 modules for stages 2 and 3.

An MST can be based on an IRT model for latent trait estimation, module assembly method and Routing.

Figure 3 – Diagram of an example of a MST with one panel and three stages.



The implementation of an IRT-MST can be separated in two parts: the assembly and the application of the MST.

To build an MST, set of items are assembled as modules following some criteria. Then, panels are constructed using those modules. It is also necessary to specify the Routing rules.

Module assembly methods of an MST can be made by the selecting set of items that maximizes or surpasses thresholds values of some elected information measurement for fixed θ's. The item set may also need to satisfy all test constraints. The most common information measurement used in MST is the Fisher. However, in this work, we will present other 2 approaches: the Kullback-Leibler information (KL) and the Continuous Entropy Method (CEM).

To assemble panels, Automated Test Assembly algorithms are available. However, since in this dissertation it will be implemented only one panel, this step will not be approached in here.

In relation to the Routing, several approaches can be implemented, including: the module selection according to the number of correct responses of an individual or according to cut off points for the θ estimates. These cut off points can be based on information functions or according to the latent trait distribution.

For the application of MST, an estimation method for the latent trait parameters, that can be done by the usual methods from IRT models, and an algorithm that administers the module according to the Routing are necessary.

In this dissertation, the Routing was determined by cut off points for θ based on the information functions or module difficulty, and the latent trait estimation used was Bayesian Expected a Posteriori method.

The Fisher, KL and CEM are presented in the sections 4.1, 4.2 and 4.3, respectively.

4.1 Fisher information

The Fisher Information is widely used in item selection procedures in adaptive testing. According to Cramér-Rao bound, under certain conditions, the inverse of this function is a lower bound for a non-biased estimator's variance. For the module selection criteria in this dissertation, the method used consists in: fixing θ values that correspond to the modules' difficulty, then, a certain number of items that have the greatest Fisher information values for those points are built in a module.

For dichotomous IRT models, the Fisher Information for one item is defined by

$$\begin{aligned} F(\theta) &= -E\left[\frac{\partial^2 \log P_{IRT}(X \mid \theta, \boldsymbol{\zeta})}{\partial \theta^2}\right] \\ &= \frac{(P'_{IRT}(X \mid \theta, \boldsymbol{\zeta}))^2}{P_{IRT}(X \mid \theta, \boldsymbol{\zeta})[1 - P_{IRT}(X \mid \theta, \boldsymbol{\zeta})]} \end{aligned} \tag{4.1}$$

where X is the examinee response to the item, $P_{IRT}(X \mid \theta, \boldsymbol{\zeta})$ is a dichotomous IRT model, θ is the examinee latent trait and $\boldsymbol{\zeta}$ is the vector of the model parameters.

The LPE Fisher information for one item is written as

$$F(\theta) = \frac{[a(1-c)\lambda p(\theta)^{\lambda}(1-p(\theta))]^2}{P(X \mid \theta, \boldsymbol{\zeta})(1 - P(X \mid \theta, \boldsymbol{\zeta}))}, \tag{4.2}$$

where $p(\theta) = \frac{1}{1+\exp(-a(\theta-b))}$ (the 2PL model) and $P(X \mid \theta, \boldsymbol{\zeta})$ defined in (2.1) with index i and j suppressed. For the construction and application of an MST the item parameters are considered known.

4.2 Kullback-Leibler information

The Kullback-Leibler (KL) information measures the distance between two functions. The greater distance the greater information value. Let θ and θ_0 be the real and fixed value of the latent trait, respectively. For one item the KL information is defined as

$$KL(\theta \mid\mid \theta_0) = E\left[\log \frac{P_{IRT}(X \mid \theta, \boldsymbol{\zeta})}{P_{IRT}(X \mid \theta_0, \boldsymbol{\zeta})}\right],$$

where $P_{IRT}(X \mid \theta, \boldsymbol{\zeta})$ is defined in (4.1). Since the true latent trait value (θ) is unknown, Sands e Waters (1996) proposed to integrate (4.3) in θ from $[\theta_0 - \delta, \theta_0 + \delta]$, where δ assumes a

value of a positive function that decreases as more items are administered in the test. Follow this suggestion, the KL information under the LPE model for one item based is given by

$$KL(\theta \mid\mid \theta_0) = \int_{\theta_0-\delta}^{\theta_0+\delta} \left[p_2(\theta) \log\left(\frac{p_2(\theta)}{p_2(\theta_0)}\right) + (1-p_2(\theta)), \log\left(\frac{1-p_2(\theta)}{1-p_2(\theta_0)}\right)\right] d\theta,$$

where $p_2(\theta) = P(X \mid \theta, \boldsymbol{\zeta})$ defined in (4.2). Chang e Ying (1996) recommended that $\delta = \frac{r}{\sqrt{k}}$, where k is the number of items administered so far and r a constant.

4.3 Continuous Entropy Method

In his work about methods of modulation, Shannon (1984) proposed the Shannon Entropy, a measurement for the uncertainty of an discrete space. The shannon entropy is defined by

$$SH(\mathbf{p}) = -C \sum_{s=1}^{S(Y)} p_s \log(p_s),$$

where, Y is a discrete random variable, $\mathbf{p} = \{p_1, ..., p_{S(Y)}\}$, C is positive constant, and $S(Y)$ are the number of all possible states of Y.

The Continuous Entropy (CE) method Wang e Chang (2011) is an adaptation of Shannon Entropy for continuous random variables. Let $\mathbf{X}_{k-1} = \{X_1, ..., X_{k-1}\}$ be a vector of the responses $X_j, j \in \{1, ..., k-1\}$, after $k-1$ items administered in the test. The response X_i assumes value 1 if the examinee chose the correct response for the item i, otherwise it assumes value 0. The Continuous Entropy function for LPE model after $k-1$ items administered in the test is written as

$$CE(\theta \mid \mathbf{X}_{k-1}) = -\int \pi(\theta \mid \mathbf{X}_{k-1}) \log(\pi(\theta \mid \mathbf{X}_{k-1})) d\theta,$$

where $\pi(\theta \mid \mathbf{X}_{k-1}) \propto \prod_{i=1}^{k-1} p_2^{X_j} (1-p_2)^{1-X_j} f(\theta)$ is the posterior distribution and $f(\theta)$ is the priori of θ. In this study, it is assumed that $\theta \sim N(\mu, \sigma^2)$

CE reaches its minimum value when the distribution $\pi(\theta \mid \mathbf{X}_{k-1})$ is concentrated on a single point i.e., $\pi(\theta = \theta_0) = 1$ and $\pi(\theta \neq \theta_0) = 0$ and its maximum value when θ has a uniform distribution.

Since the individual response for the kth item (X_k) is unknown, it is necessary to use the expected posterior continuous entropy (ECE) which is written as

$$ECE_k = -\sum_{x=0}^{1} \int \pi(\theta \mid \mathbf{X}_{k-1}, X_k = x) \log(\pi(\theta \mid \mathbf{X}_{k-1}, X_k = x)) P(X_k = x \mid \mathbf{X}_{k-1}) d\theta,$$

where $P(X_k = x \mid \mathbf{X}_{k-1})$ is the posterior predictive distribution.

4.4 Information/Entropy and LPE Item

In this section, the Fisher and KL informations and CEM under the LPE model were analyzed. To visualize the influence of the model parameters, the information/entropy of one item is considered each time.

Figures 4 and 5 present the Fisher and Kullback-Leibler Informations and Continuous Entropy Method versus the b parameter. The other parameters $(a, \theta$ and $\lambda)$ will be assigned with fixed values in these graphics.

For each function, a graphics is shown with three different information curves for three different values of $\lambda \in \{0.5, 1, 2\}$ and a single value for $a = 1$ and $\theta = 0$,

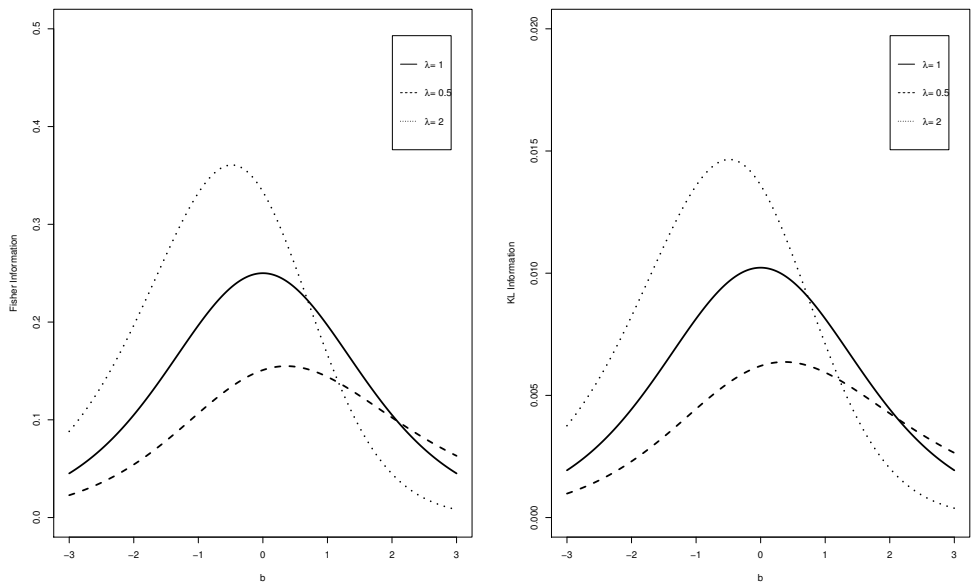


Figure 4 – Fisher and Kullback-Leibler Informations vs b.

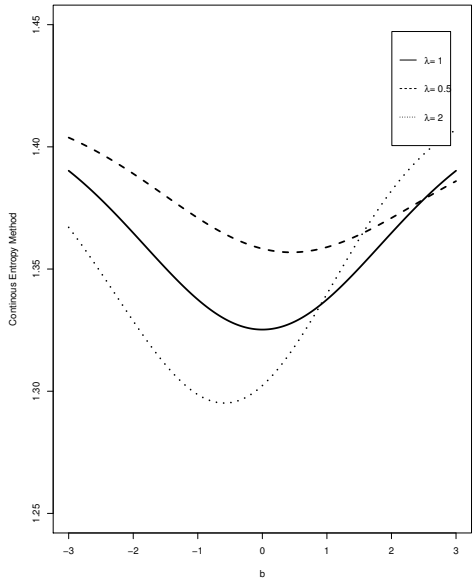


Figure 5 – Continuous Entropy Method vs b.

The Figures 4 and 5 show that the three function favor items that are not at 50% correct response for the adopted asymmetric models. For $\lambda = 0.5$, items with highest function values (or lowest ECE values) with probability of positive response of 0.6410, 0.6353 and 0.6240 for F, KL and CEM, respectively, for $\lambda = 2$, items with highest function values (or lowest ECE values) with probability of positve response 0.3816, 0.3904 and 0.4139 for F, KL and CE, respectively. It is worthy note that the CE method would selected item is closer to 50% than the others.

CHAPTER

5

POSTERIOR PREDICTIVE MODEL-CHECK AND INFORMATION STATISTICS

An item fit analysis based on the Posterior Predictive Model-Check is introduced in this chapter. This method, suggested by Rubin (1984), consists in replicating the data and comparing them with the observed data.

In the MCMC context, this method can be executed by drawing N^{fit} set of values from the model using the instances from the chains generated for parameter estimation. Then, each of these sets are used in the likelihood distribution to simulated a new dataset. The observed and expected frequency for each possible score of each item were found for the original and replicate data, respectively. The goodness of item fit can be verified by plotting the frequency of the observed score and the empirical credibility interval of the generated data.

To not rely just in a graphical approach, the PPMC allows the calculation of discrepancy measures $D(x,\zeta)$ Gelman Xiao-Li Meng (1996) using observed and replicated frequency proportions (empirical probabilities of correct responses). Then, a comparison of their posteriori distributions can be made. A significantly difference between these distributions indicates that the model didn't fit well the data.

One of the discrepancy measure proposed by in Sinharay (2006) article was the $S-X^2$ item fit statistic. This statistic can be written as

$$S-X_i^2 = \sum_{k=1}^{I-1} N_k \frac{(\rho_{ik}-E_{ik})^2}{E_{ik}(1-E_{ik})},$$

where, $k \in \{0,\ldots,I\}$ is all possible total score of the test; $i \in \{1,\ldots,I\}$ is the item index; ρ_{ik} and E_{ik} are the empirical probability (considering the observed or replicates data) and expected proportion of correct response for item i considering only the data of individuals with total score equals to k, respectively; notice that $\rho_{i0} = E_{i0} = 0$, $\rho_{iI} = E_{iI} = 1$ and $S-X_0^2 = S-X_I^2 = 0$; and

N_k is the number of individuals with score k.

The expected proportion of correct response is given by

$$E_{ik} = \frac{\int P(X=1 \mid \theta_{n^{fit}}, \zeta_{n^{fit}}) S_{k-1}(\zeta_{n^{fit}})^{*i} f(\theta) d\theta}{\int S_k(\theta_{n^{fit}}, \zeta_{n^{fit}}) f(\theta) d\theta},$$

where $n^{fit} \in \{1, ..., N^{fit}\}$ are index of samples of the MCMC chains, $\theta_{n^{fit}}$ and $\zeta_{n^{fit}}$ are the set of the n^{fit}-th selected value of the chain of individual and item parameters, respectively; $P(X=1 \mid \theta_{n^{fit}}, \zeta_{n^{fit}})$ is the probability of correct response given the current individual and item parameters; $S_k(\theta_{n^{fit}}, \zeta_{n^{fit}})$ is the probability of the score k and $S_{k-1}(\theta_{n^{fit}}, \zeta_{n^{fit}})^{*i}$ is the probability of score $k-1$ excluding item i from the pool. S_k can be calculated using a recursive approach (Lord and Wingersky, 1984).

Then, the posteriori predictive p-value (ppp-value), analogous to the p-value in a frequentist approach, can be calculated as

$$\begin{aligned} ppp-value &= P(D(X^{rep}, \theta_{n^{fit}}, \zeta_{n^{fit}})) \geq P(D(X, \theta_{n^{fit}}, \zeta_{n^{fit}})) \\ &= \int_{\theta_{n^{fit}}, \zeta_{n^{fit}}} \int_{X^{rep}} \left[I_{[D(X^{rep}, \theta_{n^{fit}}, \zeta_{n^{fit}}) \geq D(X, \theta_{n^{fit}}, \zeta_{n^{fit}})]} P(X^{rep} \mid \theta_{n^{fit}}, \zeta_{n^{fit}}) \right. \\ &\quad \left. \pi(\theta_{n^{fit}}, \zeta_{n^{fit}} \mid X) dX^{rep} d\theta_{n^{fit}}, \zeta_{n^{fit}} \right], \end{aligned} \tag{5.1}$$

where, $P(D(X^{rep}, \theta_{n^{fit}}, \zeta_{n^{fit}}))$ and $P(D(X, \theta_{n^{fit}}, \zeta_{n^{fit}}))$ are the replicates and observed discrepancy posteriori distribution, respectively; $I_{[.]}$ indicator function; $P(X^{rep} \mid \theta_{n^{fit}}, \zeta_{n^{fit}})$ is the likelihood function and $\pi(\theta_{n^{fit}}, \zeta_{n^{fit}} \mid X)$ is the posteriori function.

Since is hard to calculate ppp-value, an estimation can be done by calculating the proportion of $D(x^{rep}, \theta_{n^{fit}}, \zeta_{n^{fit}})$ were greater than $D(x, \theta_{n^{fit}}, \zeta_{n^{fit}})$. Because the null hypotheses distribution of $S - X^2$ is χ^2(I-4), a ppp-value close to zero (for example: $ppp-value$<0.05) is an indicative that the model didn't fit well the data.

Another model comparison method used in this article are based on information statistics such as: the Expected Akaike, Bayesian and Deviance information criteria.

The E-AIC, E-BIC and DIC functions can be estimated as shown in equations (5.2), (5.3) and (5.4), respectively.

$$\begin{aligned} \bar{D} &= -\frac{2}{N^C} \left(\sum_{n^c=1}^{N^C} \log L(\boldsymbol{X} \mid \theta_{n^c}, \zeta_{n^c}) \right) \\ D(\bar{\theta}_{n^c}, \bar{\zeta}_{n^c}) &= -2 \left(\log L(\boldsymbol{X} \mid \bar{\theta}_{n^c}, \bar{\zeta}_{n^c}) \right) \\ P_D &= \bar{D} - D(\bar{\theta}_{n^c}, \bar{\zeta}_{n^c}) \\ \widehat{EAIC} &= \bar{D} + 2p \qquad (5.2) \\ \widehat{EBIC} &= \bar{D} + p \log(n) \qquad (5.3) \\ \widehat{DIC} &= \bar{D} + P_D \qquad (5.4) \end{aligned}$$

where, $n^c \in \{1, ..., N^C\}$ are index of samples of the MCMC chains, $\theta_{n^c}, \zeta_{n^c}$ are the item and individual parameter values of n^c-th sample, $\bar{\theta}_{n^c}, \bar{\zeta}_{n^c}$ is the mean of these values, p is the number of parameters in the model and n is the number of individuals.

CHAPTER

6

LPE RESULTS

In this chapter, three studies of the LPE will be presented. The first two used simulated data and the third used a real data set. First, section 6.1 presents a comparison of the latent trait EAP estimates for “Rasch" LPE, 2PL and 3PL to better understand the “Rasch" LPE model. In section 6.2, The Marginal Maximum Likelihood and the Bayesian approach using MCMC methods for the item calibration process were implemented to analyze the item parameter recovery of the “Rasch" LPE model. For this study two generated sample size were used: one with 1000 and the other with 5000 individuals. In section 6.3, “Rasch" LPE was used to fit a real data set from a general science knowledge test. The verification of the goodness of fit was also implemented in this section.

6.1 LPE Model and its θ estimate characteristics

To analyze the differences on latent trait estimates between LPE’s and other logistic models, two schemes were simulated and the results are presented in this section.

For both schemes, the data was generated considering 5 items and 32 examinee responses, each examenee corresponding to a different response pattern and all possible combinations were contemplated. The item parameters were considered known and the EAP method was used to estimate the latent traits for each of the models.

The first scheme were design to compare the latent trait estimate results from the Rasch, “Rasch" LPE and 3PL models when inconsistent response patterns are given. For this analysis, the b parameter values of all models were set as: -3, -1.5, 0, 1.5 and 3. For the “Rasch LPE", the acceleration parameter values were fixed at $\lambda = 0.5$ or $\lambda = 2$ for all items. For 3 Parameter Logistic (3PL) model, $c = 0.2$ was considered for all items.

In Table 1, the θ estimates for each model and each response pattern are presented:

Table 1 – Rasch, “Rasch" LPE and 3PL θ estimates for each response pattern

Responses	Rasch	“Rasch" LPE $\lambda = 0.5$	“Rasch" LPE $\lambda = 2$	3PL $c = 0.2$
0,0,0,0,0	-1.55	-1.798	-1.281	-1.55
0,0,0,0,1	-0.92	-1.444	-0.101	-1.506
0,0,0,1,0	-0.92	-1.41	-0.22	-1.40
0,0,1,0,0	-0.92	-1.33	-0.39	-1.18
0,1,0,0,0	-0.92	-1.25	-0.54	-1.01
1,0,0,0,0	-0.92	-1.194	-0.62	-0.94
0,0,0,1,1	-0.31	-1.03	0.76	-1.34
0,0,1,0,1	-0.31	-0.95	0.61	-1.11
0,0,1,1,0	-0.31	-0.90	0.47	-0.95
0,1,0,0,1	-0.31	-0.87	0.52	-0.93
1,0,0,0,1	-0.31	-0.82	0.48	-0.87
0,1,0,1,0	-0.31	-0.82	0.38	-0.77
1,0,0,1,0	-0.31	-0.77	0.34	-0.71
0,1,1,0,0	-0.31	-0.74	0.21	-0.54
1,0,1,0,0	-0.31	-0.69	0.17	-0.49
1,1,0,0,0	-0.31	-0.61	0.05	-0.37
0,0,1,1,1	0.31	-0.49	1.32	-0.84
0,1,0,1,1	0.31	-0.41	1.27	-0.66
1,0,0,1,1	0.31	-0.37	1.26	-0.60
0,1,1,0,1	0.31	-0.32	1.14	-0.42
1,0,1,0,1	0.31	-0.29	1.12	-0.37
0,1,1,1,0	0.31	-0.25	0.99	-0.20
1,0,1,1,0	0.31	-0.22	0.97	-0.15
1,1,0,0,1	0.31	-0.21	1.06	-0.24
1,1,0,1,0	0.31	-0.15	0.90	-0.04
1,1,1,0,0	0.31	-0.06	0.75	0.16
0,1,1,1,1	0.92	0.20	1.79	-0.01
1,0,1,1,1	0.92	0.23	1.79	0.04
1,1,0,1,1	0.92	0.29	1.75	0.15
1,1,1,0,1	0.92	0.38	1.65	0.36
1,1,1,1,0	0.92	0.46	1.49	0.60
1,1,1,1,1	1.55	0.95	2.27	0.91
b parameter values -3,-1.5,0,1.5,3				

In Table 2, the rank (greater ranks means greater θ estimates) of the responses patterns are presented.

Notice in Table 1 and 2 that, for the LPE with $\lambda = 0.5$ the higher number of correct responses the higher θ estimates and rank, but for different values of $\lambda < 1$ and b this relationship may not hold.

The patterns 1,1,1,1,0 and 0,1,1,1,1 points out different aspects of each model. The former portrays a consistent response pattern and the latter portrays the case of an individual with high proficiency level that accidentally misses one easy item. The LPE with $\lambda = 0.5$ and the 3PL penalizes more the wrong answer for the easiest item, $\theta_{11110} > \theta_{01111}$. The LPE with $\lambda = 2$ rewards more the right answer for the most difficult item, $\theta_{11110} < \theta_{01111}$, meaning the mistake on the easy item will not have much impact in the estimate.

Another interesting comparison is between the patterns 0,0,0,0,1 and 1,0,0,0,0. As in the previous case, the LPE with $\lambda = 0.5$ and 3PL penalizes more the wrong answer for the easiest item, $\theta_{10000} > \theta_{00001}$, and the LPE with $\lambda = 2$ rewards more the right answer for the most difficult item, $\theta_{10000} < \theta_{00001}$. However, the former models are “predicting" that the individual is probably guessing the most difficult item, and not rewarding them much, but the LPE with

Table 2 – Rasch, “Rasch" LPE and 3PL ranks for each response pattern

Responses	Rasch	“Rasch" LPE $\lambda = 0.5$	“Rasch" LPE $\lambda = 2$	3PL $c = 0.2$
0,0,0,0,0	1	1	1	1
0,0,0,0,1	4	2	6	2
0,0,0,1,0	4	3	5	3
0,0,1,0,0	4	4	4	5
0,1,0,0,0	4	5	3	7
1,0,0,0,0	4	6	2	9
0,0,0,1,1	11.5	7	17	4
0,0,1,0,1	11.5	8	15	6
0,0,1,1,0	11.5	9	12	8
0,1,0,0,1	11.5	10	14	10
1,0,0,0,1	11.5	11	13	11
0,1,0,1,0	11.5	12	11	13
1,0,0,1,0	11.5	13	10	14
0,1,1,0,0	11.5	14	9	17
1,0,1,0,0	11.5	15	8	18
1,1,0,0,0	11.5	16	7	21
0,0,1,1,1	21.5	17	26	12
0,1,0,1,1	21.5	18	25	15
1,0,0,1,1	21.5	19	24	16
0,1,1,0,1	21.5	20	23	19
1,0,1,0,1	21.5	21	22	20
0,1,1,1,0	21.5	22	20	23
1,0,1,1,0	21.5	23	19	24
1,1,0,0,1	21.5	24	21	22
1,1,0,1,0	21.5	25	18	25
1,1,1,0,0	21.5	26	16	29
0,1,1,1,1	29	27	31	26
1,0,1,1,1	29	28	30	27
1,1,0,1,1	29	29	29	28
1,1,1,0,1	29	30	28	30
1,1,1,1,0	29	31	27	31
1,1,1,1,1	32	32	32	32
b parameter values -3,-1.5,0,1.5,3				

$\lambda = 2$ is giving him bigger score.

The Pearson correlations of the ranks between each model is shown in the Table 3. It shows that the ranks between models are highly correlated, meaning that the latent traits produced by each model were ordered in similar fashion. The LPE’s with $\lambda = 0.5$ and $\lambda = 2$ were the least similar and the LPE with $\lambda = 0.5$ and 3PL with $c = 0.2$ had the most similar ranks.

Table 3 – Correlation of the ranks between the IRT models

	Rasch	“Rasch" LPE $\lambda = 0.5$	“Rasch" LPE $\lambda = 2$	3PL $c = 0 - .2$
Rasch	1			
“Rasch" LPE $\lambda = 0.5$	0.9655	1		
“Rasch" LPE $\lambda = 2$	0.8596	0.6261	1	
3PL $c = 0.2$	0.8905	0.9751	0.7360	1

In fact, in general, the LPE and 3PL ordered the response patterns like the Rasch model, but with and imbued tie criteria. LPE with $\lambda = 0.5$ and 3PL with $c = 0.2$ favors the easier items and LPE’s $\lambda = 2$ favors the more difficult ones.

In the second scheme, two comparisons between the 2PL and ”Rasch" LPE is shown

in Table 4. In the first one, the 2PL a parameter were 1.5, 1, 0.5, 1, 1.5 for the items with b parameters -3, -1.5, 0, 1.5, 3, respectively, and for the "Rasch" LPE the λ values are 2, 1, 0.5, 1, 2 respectively. In the second comparison, the 2PL a parameter were 0.5, 0.5, 1, 1.5, 1.5, respectively, and for the "Rasch"LPE the λ values are 0.5, 0.5, 1, 2, 2 respectively. This design was made to compare the impact of the discrimination (a) and acceleration (λ) parameters on the latent trait estimates.

Table 4 – "Rasch" LPE and 2PL ranks for each response pattern

Responses	Comparison 1		Comparison 2	
	2PL	"Rasch"LPE	2PL	"Rasch"LPE
	a=1.5,1,0.5,1,1.5	λ = 2,1,0.5,1,2	a=0.5,0.5,1,1.5,1.5	λ = 0.5,0.5,1,2,2
0,0,0,0,0	1.0	1.0	1.0	1.0
1,0,0,0,0	6.5	5.0	2.5	3.0
0,1,0,0,0	3.5	3.5	2.5	2.0
0,0,1,0,0	2.0	2.0	4.5	4.0
0,0,0,1,0	3.5	3.5	7.5	5.0
0,0,0,0,1	6.5	9.0	7.5	7.0
1,1,0,0,0	14.0	11.5	4.5	6.0
1,0,1,0,0	10.0	8.0	7.5	9.0
0,1,1,0,0	6.5	6.5	7.5	8.0
1,0,0,1,0	14.0	11.5	12.0	11.0
0,1,0,1,0	10.0	10.0	12.0	10.0
0,0,1,1,0	6.5	6.5	16.5	12.0
1,0,0,0,1	19.0	17.0	12.0	14.0
0,1,0,0,1	14.0	15.5	12.0	13.0
0,0,1,0,1	10.0	13.0	16.5	15.0
0,0,0,1,1	14.0	15.5	21.0	17.0
1,1,1,0,0	19.0	18.5	12.0	16.0
1,1,0,1,0	23.0	20.0	16.5	18.0
1,0,1,1,0	19.0	18.5	21.0	20.0
0,1,1,1,0	14.0	14.0	21.0	19.0
1,1,0,0,1	26.5	25.5	16.5	21.0
1,0,1,0,1	23.0	23.0	21.0	23.0
0,1,1,0,1	19.0	21.5	21.0	22.0
1,0,0,1,1	26.5	25.5	25.5	25.0
0,1,0,1,1	23.0	24.0	25.5	24.0
0,0,1,1,1	19.0	21.5	28.5	26.0
1,1,1,1,0	26.5	27.0	25.5	27.0
1,1,1,0,1	29.5	29.5	25.5	28.0
1,1,0,1,1	31.0	31.0	28.5	29.0
1,0,1,1,1	29.5	29.5	30.5	31.0
0,1,1,1,1	26.5	28.0	30.5	30.0
1,1,1,1,1	32.0	32.0	32.0	32.0
b parameter values -3,-1.5,0,1.5,3				

The correlation between the ranks in comparison 1 is 0.987 and in comparison 2 is 0.977. In comparison 1, for the response pattern 1,0,0,0,0 and 0,0,0,0,1, the 2PL model produced the same rank. The "Rasch" LPE model favored the latter due to the λ values for those items being greater than 1. This same effect can be seen in few others responses patterns.

The pattern 0,1,1,1,1 and 1,1,1,1,0 of the "Rasch LPE" in case 2 evidences that even though there are items with λ lower than 1, the latter pattern has higher rank.

6.2 MCMC vs MML for the LPE model

To study the quality of recovery, simulation of the "*Rasch*" LPE model considering two individual sample sizes were made with latent traits generated from $Normal(0,1)$: the first one, with 1000 individuals, recommend for simpler IRT models, and the second one, with 5000 individuals, because this sample size produced relative well latent trait recovery in a feasible amount of time in the present study. Pondering about the time consumed by the estimation algorithms and the usual simulated item pool size in literature, 20 items based on the "Rasch" LPE model were genarated for each case. The $Normal(0,1)$ were used to generate the b's values and the values 0.5, 0.6, 0.7, 0.8, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1.5 and 2 were fixed for λ's. Notice that 20% of the values are < 1 and the other 10% are > 1. These values were based on Flores (2012) and Bolfarine e Bazan (2010b).

The MCMC method was implemented via Winbugs, and the Maximum Marginal Likelihood (MML) method was implemented in R.

For the MCMC, 10 replicates were made with prior distributions for the parameter as: $Normal(0,10)$ for b_j, $gamma(0.25, rate = 0.25)$ for λ_j; for $j = \{1,...,20\}$, n is the number of individuals and $Normal(0,1)$ for θ_i, for $i = \{1,...,n\}$. Three chains with random initial values, 50000 iterations, 20000 burn-in and thinning of 10 were considered.

To evaluate the precision of the estimates results for the replicates, some measurements based on the true and estimated values were calculated as presented below.

$$BIAS = \frac{\sum_{i=1}^{n}(\theta_i - \hat{\theta}_i)}{n}, \tag{6.1}$$

where θ_i is the real latent trait, $\hat{\theta}_i$ is the latent trait estimate and n is the number of individuals, and

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\theta_i - \hat{\theta}_i)^2}{n}}. \tag{6.2}$$

For MML simulation study, the data of one of these replicates was used. No replicates were made for the MML because of the time required.

Case 1: 1000 individuals

In Tables 5 and 6, the item parameter true values, MML and MCMC estimates using the same generated data (MML_1 and $MCMC_1$, respectively), mean of MCMC replicate estimates, MCMC credibility interval (CI), MCMC bias and RMSE for the b and λ parameters, respectively, are presented. In the last row, the absolute difference average between the true and estimated is shown.

Table 5 – Simulated results for b's parameters for 1000 individuals with credibility interval (CI) and average absolute difference between true and estimated values (Avg. $|\varsigma - \hat{\varsigma}|$).

			10 replicates with MCMC method				
True values	MML_1	$MCMC_1$	Means	CI		Bias	RMSE
-0.43	-0.61	-0.56	-0.54	(-1.23	, 0.44)	-0.11	0.59
0.60	1.07	0.92	0.45	(0.06	, 0.93)	-0.15	0.32
0.12	0.50	0.45	0.06	(-0.58	, 0.44)	-0.06	0.34
-1.16	-0.90	-0.83	-0.75	(-1.17	, -0.18)	0.41	0.54
0.82	0.60	0.48	0.61	(0.02	, 1.34)	-0.21	0.47
-0.11	1.04	0.91	0.13	(-0.80	, 0.88)	0.24	0.61
-0.45	-0.12	-0.20	-0.47	(-0.99	, 0.21)	-0.02	0.40
-2.41	-3.06	-1.86	-1.30	(-1.89	, -0.25)	1.11	1.22
-2.43	-3.35	-1.81	-1.28	(-2.09	, -0.44)	1.15	1.31
1.98	2.41	2.08	1.54	(1.01	, 2.09)	-0.44	0.62
-1.31	-0.32	-0.32	-0.93	(-1.71	, -0.35)	0.38	0.58
1.02	0.99	0.87	0.68	(0.02	, 1.08)	-0.34	0.47
-1.63	-1.03	-0.85	-0.99	(-1.82	, -0.46)	0.64	0.76
-0.51	-0.93	-0.91	-0.61	(-0.98	, 0.24)	-0.10	0.40
-0.21	-0.26	-0.30	-0.37	(-0.79	, 0.29)	-0.16	0.36
0.62	1.11	1.02	0.47	(-0.36	, 0.98)	-0.16	0.45
-0.49	-1.08	-1.01	-0.58	(-1.16	, 0.16)	-0.09	0.42
-0.65	-0.93	-0.83	-0.60	(-1.24	, -0.05)	0.05	0.40
-0.80	-0.69	-0.73	-0.80	(-1.36	, -0.36)	0.00	0.32
0.87	0.13	0.09	0.50	(-0.01	, 1.55)	-0.37	0.61
Avg. $\lvert\varsigma - \hat{\varsigma}\rvert$	0.46	0.42	0.31				

Table 6 – Simulated results for λ's parameters for 1000 individuals with credibility interval (CI) and average absolute difference between true and estimated values (Avg. $|\varsigma - \hat{\varsigma}|$).

			10 replicates with MCMC method				
True values	MML_1	$MCMC_1$	Means	CI		Bias	RMSE
0.5	0.62	0.68	0.70	(0.28	, 1.16)	0.20	0.36
0.6	0.47	0.53	0.71	(0.53	, 0.91)	0.11	0.16
0.7	0.58	0.63	0.80	(0.63	, 1.21)	0.10	0.22
0.8	0.72	0.79	0.70	(0.43	, 0.96)	-0.10	0.22
1.0	1.16	1.31	1.26	(0.81	, 1.82)	0.26	0.42
1.0	0.46	0.51	1.00	(0.52	, 1.89)	-0.00	0.44
1.0	0.88	1.00	1.18	(0.69	, 1.65)	0.18	0.38
1.0	2.17	0.90	0.55	(0.23	, 0.93)	-0.45	0.50
1.0	2.61	0.80	0.54	(0.24	, 0.93)	-0.46	0.53
1.0	0.84	1.00	1.38	(0.96	, 1.80)	0.38	0.51
1.0	0.47	0.52	0.94	(0.53	, 1.67)	-0.06	0.38
1.0	0.97	1.07	1.36	(1.02	, 2.26)	0.36	0.54
1.0	0.70	0.70	0.77	(0.43	, 1.55)	-0.23	0.42
1.0	1.51	1.64	1.24	(0.62	, 1.70)	0.24	0.41
1.0	1.10	1.22	1.30	(0.83	, 1.75)	0.30	0.40
1.0	0.79	0.85	1.28	(0.87	, 2.35)	0.28	0.56
1.0	1.66	1.78	1.27	(0.66	, 2.09)	0.27	0.50
1.0	1.25	1.27	1.14	(0.69	, 1.94)	0.14	0.42
1.5	1.38	1.59	1.71	(1.14	, 2.76)	0.21	0.56
2.0	3.41	3.80	2.95	(1.42	, 4.01)	0.95	1.24
Avg. $\lvert\varsigma - \hat{\varsigma}\rvert$	0.42	0.31	0.26				

Tables 5 and 6 show that the parameter estimates obtained with a particular sample of 1000 individual responses to 20 items under the “Rasch" LPE model were closer to their true values when using MCMC than MML.

In the MCMC results, 90% the b parameters' credibility intervals contained the true values of the parameter. while for λ, this occurred in 80% of the credibility intervals. The amplitude of the CIs is very high, which means that the estimates generated by MCMC method

had a great variance.

Notice that, if the CIs were used as an evidence to verify if an item is asymmetric or not, two of the $\lambda < 1$ would be considered as symmetric and two of the symmetric items would be considered asymmetric.

For $\lambda = 2$ the Bias and RMSE were large, showing that this parameter were poorly recovered for MCMC. The MML estimate was also very far from the real value.

The correlations, in this case, between the b and λ estimates are -0.64 and -0.51 for the MCMC and MML methods, respectively and the true correlation is about 0.10. This results indicates that for a small sample, there might be an identifiability problem.

Case 2: 5000 individual

Tables 7 and 8 present the item parameter true values, MML_1 (iteration 200 and maximum difference of 0.0012 between iterations of parameter estimates) and MCMC_1 estimates, mean of MCMC replicate estimates, MCMC credibility interval, MCMC bias and RMSE for the b and λ parameters, respectively. The absolute difference average between the true and estimated values are displayed in the last row of the table.

Table 7 – Simulated results for b's parameter for 5000 individuals with credibility interval (CI) and average absolute difference between true and estimated values (Avg. $|\varsigma - \hat{\varsigma}|$).

			10 replicates with MCMC method			
True values	MML_1	MCMC_1	Means	CI	Bias	RMSE
-0.43	-0.58	-0.56	-0.39	(-0.62 , 0.00)	0.04	0.19
0.60	0.31	0.31	0.57	(0.25 , 0.86)	-0.03	0.20
0.12	0.43	0.43	0.20	(-0.15 , 0.56)	0.08	0.23
-1.16	-1.08	-0.98	-1.02	(-1.39 , -0.56)	0.14	0.28
0.82	0.88	0.87	0.79	(0.64 , 0.91)	-0.03	0.10
-0.11	0.25	0.26	-0.02	(-0.30 , 0.24)	0.09	0.20
-0.45	-0.18	-0.16	-0.30	(-0.70 , 0.15)	0.15	0.31
-2.41	-2.13	-1.77	-1.92	(-2.70 , -1.24)	0.49	0.64
-2.43	-2.18	-1.75	-2.07	(-2.48 , -1.66)	0.36	0.45
1.98	1.81	1.76	1.92	(1.61 , 2.17)	-0.06	0.19
-1.31	-1.10	-1.05	-1.20	(-1.63 , -0.75)	0.11	0.31
1.02	1.29	1.26	1.00	(0.75 , 1.24)	-0.02	0.16
-1.63	-1.22	-1.11	-1.38	(-1.89 , -1.05)	0.25	0.39
-0.51	-0.71	-0.68	-0.44	(-0.67 , -0.20)	0.07	0.17
-0.21	-0.11	-0.10	-0.17	(-0.54 , 0.09)	0.04	0.19
0.62	0.40	0.41	0.70	(0.42 , 1.03)	0.08	0.22
-0.49	-0.50	-0.46	-0.55	(-0.99 , -0.14)	-0.06	0.30
-0.65	-0.58	-0.59	-0.55	(-1.03 , -0.27)	0.10	0.28
-0.80	-0.83	-0.87	-0.69	(-0.97 , -0.29)	0.11	0.27
0.87	0.86	0.83	0.83	(0.52 , 1.08)	-0.04	0.19
Avg. $\lvert\varsigma - \hat{\varsigma}\rvert$	0.19	0.24	0.12			

Tables 7 and 8 shown, unlike to the 1000 individual case, the MML estimator produced better results than MCMC when considering 5000 individual and 20 items, indicating that although the MML needed greater sample size, it had a better performance.

In the MCMC results, all CIs contained the true value of the parameter.

Notice that, if the CI were used as an evidence to verify if an item is asymmetric or not, there are two item that would be considered symmetric even though, they are asymmetric

Table 8 – Simulated results for λ's parameters for 5000 individuals with credibility interval (CI) and average absolute difference between true and estimated values (Avg. $|\varsigma - \hat{\varsigma}|$).

			10 replicates with MCMC method			
True values	MML_1	$MCMC_1$	Means	CI	Bias	RMSE
0.5	0.54	0.54	0.50	(0.39 , 0.58)	0.00	0.06
0.6	0.71	0.72	0.63	(0.52 , 0.75)	0.03	0.08
0.7	0.60	0.61	0.68	(0.55 , 0.85)	-0.02	0.09
0.8	0.73	0.70	0.76	(0.52 , 1.03)	-0.04	0.16
1.0	0.97	0.98	1.04	(0.98 , 1.12)	0.04	0.06
1.0	0.81	0.81	0.97	(0.82 , 1.15)	-0.03	0.12
1.0	0.82	0.82	0.93	(0.66 , 1.23)	-0.07	0.19
1.0	0.77	0.65	0.79	(0.39 , 1.62)	-0.21	0.43
1.0	0.78	0.60	0.84	(0.56 , 1.22)	-0.16	0.28
1.0	1.06	1.10	1.03	(0.91 , 1.16)	0.03	0.09
1.0	0.85	0.85	0.98	(0.65 , 1.38)	-0.02	0.25
1.0	0.87	0.89	1.02	(0.90 , 1.16)	0.02	0.09
1.0	0.72	0.69	0.88	(0.64 , 1.30)	-0.13	0.27
1.0	1.11	1.11	0.98	(0.82 , 1.14)	-0.02	0.11
1.0	0.94	0.95	0.99	(0.82 , 1.28)	-0.01	0.14
1.0	1.13	1.14	0.97	(0.81 , 1.13)	-0.03	0.12
1.0	1.03	1.02	1.09	(0.80 , 1.52)	0.09	0.27
1.0	0.92	0.95	0.97	(0.74 , 1.41)	-0.03	0.22
1.5	1.58	1.68	1.44	(0.97 , 1.81)	-0.06	0.29
2.0	2.06	2.14	2.13	(1.78 , 2.67)	0.13	0.34
Avg. $\lvert\varsigma - \hat{\varsigma}\rvert$	0.12	0.14	0.06			

($\lambda = 0.8$ and $\lambda = 1.5$). Observe that neither these item were incorrectly classified in the 1000 individual simulations.

The Bias and RMSE in this case were considerably lower than in the previous one, showing that 1000 individual sample is too small to estimate the “Rasch" LPE. However, with a higher sample size the estimate results can be decent.

The correlations, in this case, between the b and λ estimates are -0.3100 and 0.2727 for the MCMC and MML methods, respectively, while the true correlation is about 0.10. This fact point out that the identifiability problem were reduced with a large number of individuals.

Another relevant information about this study is that the 5000 individual simulation required less iterations to reach the convergence threshold than the 1000's. This indicates that “Rasch" LPE requires a large number of responses for each item to reach convergence.

6.3 Real application

In this section, the General Science (GS) test data gathered by Educational Testing Service (ETS) through Amazon Turk is analyzed. Results using the 2PL, “Rasch" LPE and MCMC or MML method are discussed.

The GS test is composed by items adapted from the SLiM Instrument Rundgren *et al.* (2012) with objective to measure the general science knowledge. The test has Cronbach's α of 0.89. The dataset consists of responses to 37 dichotomous multiple choices items by 1565 individual and it was gathered for Tetralogue project Bazaldua *et al.* (2015).

In Table 9 are presented the percent of correct responses for each item.

Table 9 – Percentage of correct responses for each item in GS test application.

item 1 86.39	item 2 79.87	item 3 83.39	item 4 86.90	item 5 79.74	item 6 78.27	item 7 69.84	item 8 92.59
item 9 87.41	item 10 82.81	item 11 74.31	item 12 79.74	item 13 84.03	item 14 77.25	item 15 73.61	item 16 90.48
item 17 91.18	item 18 84.66	item 19 75.91	item 20 87.80	item 21 54.06	item 22 42.04	item 23 55.91	item 24 53.99
item 25 87.99	item 26 91.25	item 27 89.58	item 28 5.37	item 29 31.88	item 30 65.56	item 31 78.27	item 32 71.63
item 33 89.01	item 34 87.60	item 35 45.24	item 36 48.56	item 37 73.55			

The GS test is on an easy level, with an average of 73.45% of correct responses and average score of 27.18 out of 37. Item 28 is the hardest item with around 5% of correct responses and item 36 is the closest to 50% (48.56%) of correct responses.

To analyze the “Rasch" LPE model with this dataset, six cases were considered as described in Table 10. The Case 1 consisted of fitting the 2PL model to the data using the MCMC method with priors $\theta \sim Normal(0,1)$, $b \sim Normal(0,2)$ and $a \sim Log-Normal(-0.1,\sqrt{0.5})$. In Case 2, the “Rasch" LPE was used with the MML estimation method. For Case 3 through 6, the “Rasch" LPE with the Bayesian MCMC estimation was applied.

In the Bayesian approach for the “Rasch" LPE model, Cases 3 and 4 had all 37 items, while Cases 5 and 6 had only 36 items (item 28 have issues and it was removed from the pool). For Cases 3 and 5, the b parameter priors followed $Normal(0,1)$ distributions; for Cases 4 and 6, the $Normal(0,2)$ were used instead. For these four cases, θ's and λ's priors were $Normal(0,1)$ and $Gamma(0.25,0.25)$, respectively.

The MCMC methods were implemented via Winbugs software with 3 chains, 100000 iterations with 30000 burn-in and thinning of 20 were generated. The MML (in Case 2) was implemented via R.

Table 10 shows a summarized version of the simulations and Tables 11 and 12 present the resultant b and λ (a for Case 1) parameter estimates for each Case.

Table 10 – Methodology applied to the GS data

Case	model	items	method	θ distributions/priors	items priors
Case 1	2PL	37	MCMC	Normal(0,1)	$a \sim Log-Normal(-0.1,\sqrt{0.5})$, $b \sim Normal(0,2)$
Case 2	“Rasch LPE"	37	MML	Normal(0,1)	
Case 3	“Rasch LPE"	37	MCMC	Normal(0,1)	$b \sim Normal(0,1)$, $\lambda \sim Gamma(0.25,0.25)$
Case 4	“Rasch LPE"	37	MCMC	Normal(0,1)	$b \sim Normal(0,2)$, $\lambda \sim Gamma(0.25,0.25)$
Case 5	“Rasch LPE"	36	MCMC	Normal(0,1)	$b \sim Normal(0,1)$, $\lambda \sim Gamma(0.25,0.25)$
Case 6	“Rasch LPE"	36	MCMC	Normal(0,1)	$b \sim Normal(0,2)$, $\lambda \sim Gamma(0.25,0.25)$

Table 11 – b parameter estimates for the 2PL (Case 1) and “Rasch" LPE (Cases 2-6) models

Items	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
Item 1	-2.28	-1.34	-0.73	-0.81	-0.56	-0.80
Item 2	-1.65	-0.80	-0.36	-0.40	-0.37	-0.47
Item 3	-1.53	-3.10	-1.78	-2.24	-1.80	-2.23
Item 4	-1.65	-4.02	-2.37	-3.11	-2.50	-3.08
Item 5	-1.50	-0.86	-0.48	-0.61	-0.46	-0.59
Item 6	-1.49	-0.57	-0.28	-0.36	-0.25	-0.33
Item 7	-0.90	-1.12	-0.67	-0.80	-0.60	-0.78
Item 8	-2.08	-4.53	-2.55	-3.46	-2.44	-3.46
Item 9	-1.49	-4.02	-2.55	-3.21	-2.52	-3.19
Item 10	-1.39	-3.73	-2.33	-2.95	-2.27	-2.91
Item 11	-1.26	-0.59	-0.24	-0.34	-0.24	-0.32
Item 12	-1.08	-3.81	-2.54	-3.17	-2.57	-3.15
Item 13	-1.54	-3.54	-2.16	-2.68	-1.91	-2.68
Item 14	-1.61	-0.55	-0.14	-0.21	-0.21	-0.23
Item 15	-1.16	-1.72	-0.92	-1.16	-0.93	-1.16
Item 16	-1.64	-4.65	-3.04	-3.91	-3.00	-3.91
Item 17	-1.55	-4.78	-3.33	-4.15	-3.36	-4.16
Item 18	-1.30	-4.12	-2.77	-3.47	-2.73	-3.44
Item 19	-1.10	-3.22	-1.79	-2.49	-1.91	-2.46
Item 20	-1.28	-4.60	-3.39	-4.10	-3.38	-4.08
Item 21	-0.21	1.22	1.21	1.31	1.20	1.33
Item 22	0.44	1.47	1.46	1.53	1.51	1.55
Item 23	-0.39	1.76	1.65	1.83	1.61	1.83
Item 24	-0.15	-0.69	-0.40	-0.44	-0.42	-0.44
Item 25	-1.71	-3.92	-2.36	-2.96	-2.21	-2.98
Item 26	-2.03	-4.32	-2.40	-3.30	-2.51	-3.24
Item 27	-2.15	-4.16	-2.29	-3.03	-2.17	-3.05
Item 28	7.13	7.84	4.78	6.50	-	-
Item 29	0.94	1.29	1.25	1.37	1.25	1.40
Item 30	-1.22	1.83	1.60	1.89	1.59	1.92
Item 31	-2.23	1.14	1.04	1.29	1.03	1.30
Item 32	-1.11	-0.82	-0.37	-0.45	-0.37	-0.47
Item 33	-1.53	-4.46	-2.98	-3.75	-2.94	-3.75
Item 34	-1.63	-4.28	-2.64	-3.45	-2.67	-3.43
Item 35	0.33	2.06	1.87	2.11	1.88	2.13
Item 36	0.12	2.24	1.93	2.24	1.98	2.28
Item 37	-1.17	-0.65	-0.38	-0.40	-0.39	-0.39

The MML and MCMC produced different results in several cases. In the simulation with 1000 individual and 20 items presented in section 6.2, the MML had worse parameter recovery performance than the MCMC. The GS data has 1565 examenees and 37 items, this could indicate that the MML is less reliable than MCMC in these cases.

Comparing the 2PL (Case 1) with “Rasch" LPE, there appears to be a relationship between the a and λ. For most of the them, items with a estimated value above 1 had also λ values greater than one, and vice versa. This might happen because both a and λ impacts the inclination of the ICC. However, this might not occurs to all items because the λ also effects the positioning of the curve, interacting with the b values as well.

In general, the removal of item 28 from Cases 5 and 6 did not have much impact on the estimate when comparing with Cases 3 and 4, respectively. The λ values of item 13 from Cases 3 and 5 were the farthest apart, with their respected λ values of 1.25 and 0.99.

Notice that in Case 1 (2PL model), the item 28 had estimated a value of 0.41 (low discrimination power), which is the lowest of all items.

Table 12 – a (Case 1) and λ (Cases 2-6) parameter estimates.

Items	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
Item 1	0.94	0.52	0.32	0.34	0.28	0.33
Item 2	0.99	0.53	0.39	0.42	0.39	0.41
Item 3	1.39	3.30	0.92	1.40	0.96	1.37
Item 4	1.59	5.74	1.14	2.31	1.31	2.25
Item 5	1.11	0.55	0.42	0.45	0.42	0.45
Item 6	1.03	0.48	0.40	0.41	0.39	0.41
Item 7	1.16	1.15	0.81	0.89	0.78	0.88
Item 8	1.72	4.80	0.70	1.67	0.64	1.66
Item 9	2.04	5.45	1.30	2.42	1.25	2.37
Item 10	1.56	6.01	1.61	2.83	1.50	2.74
Item 11	1.00	0.61	0.48	0.51	0.49	0.50
Item 12	1.94	8.01	2.36	4.32	2.49	4.27
Item 13	1.44	4.67	1.25	1.98	0.99	1.99
Item 14	0.87	0.52	0.39	0.41	0.41	0.41
Item 15	1.07	1.72	0.83	1.00	0.85	1.00
Item 16	2.27	7.10	1.49	3.39	1.44	3.39
Item 17	3.00	7.46	1.80	3.93	1.84	3.97
Item 18	2.11	7.53	2.03	3.99	1.97	3.89
Item 19	1.37	5.83	1.52	2.88	1.71	2.81
Item 20	3.27	9.13	2.88	5.59	2.85	5.46
Item 21	0.82	0.44	0.44	0.42	0.45	0.41
Item 22	0.87	0.56	0.57	0.55	0.56	0.54
Item 23	0.65	0.32	0.34	0.31	0.35	0.31
Item 24	1.28	1.55	1.26	1.28	1.29	1.28
Item 25	1.62	4.71	1.05	1.81	0.91	1.85
Item 26	1.58	4.70	0.74	1.73	0.84	1.64
Item 27	1.26	5.02	0.82	1.68	0.75	1.72
Item 28	0.41	0.38	0.64	0.46	-	-
Item 29	0.98	0.85	0.88	0.82	0.88	0.81
Item 30	0.56	0.22	0.25	0.22	0.25	0.21
Item 31	0.63	0.18	0.18	0.16	0.19	0.16
Item 32	0.98	0.85	0.60	0.63	0.60	0.64
Item 33	2.26	7.01	1.65	3.46	1.63	3.43
Item 34	1.72	6.76	1.40	2.99	1.45	2.98
Item 35	0.67	0.39	0.42	0.38	0.42	0.38
Item 36	0.60	0.33	0.38	0.33	0.37	0.33
Item 37	1.05	0.66	0.55	0.55	0.55	0.55

The effects of b priors over the estimates can be seem by comparing Cases 3 and 4, or Case 5 and Case 6. Although there are no disagreements in relation to the b values signs, the magnitude differs for some items, and some examples are items 3,8, 26 and 27. In particular, these four items also had λ's < 1 or λ's > 1, depending on the prior considered. However, their 95% credibility intervals overlapped each other due to high standard deviation. More individual response samples might be required to improve the estimates.

In relation to convergence, for the MML estimation, the criteria was not reached (difference between two consecutive iterations lower than a fixed value for all items). After 150 iterations, only 35% of b's and 54% of λ's estimates reached convergence for the stopping value 0.001. Or If it was 0.005, the convergence rate would increase to 64% and 57% for b and λ, respectively. Due to the time cost of the algorithm, no more iterations were made. This fact corroborate the previous statement of MML being less reliable than MCMC. For MCMC chain convergence, Figures 6 and 7 show the German-Rubin statistic for some items from Case 4 and Case 6, respectively.

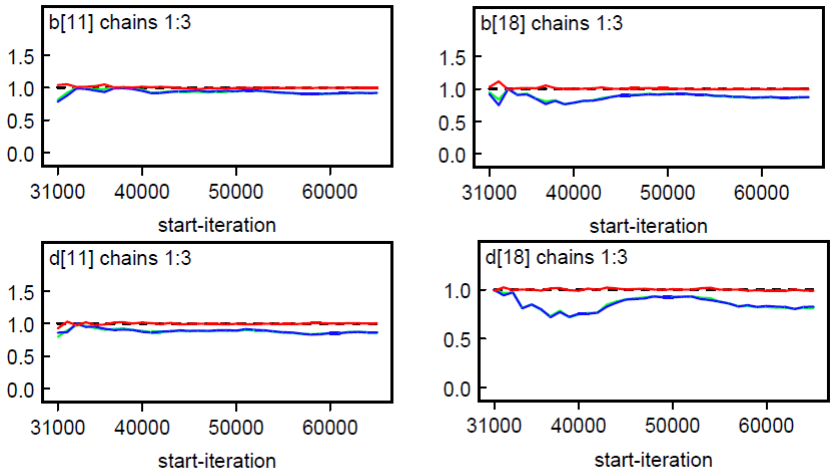


Figure 6 – Gelman Rubin convergence “Rasch" LPE for 37 items

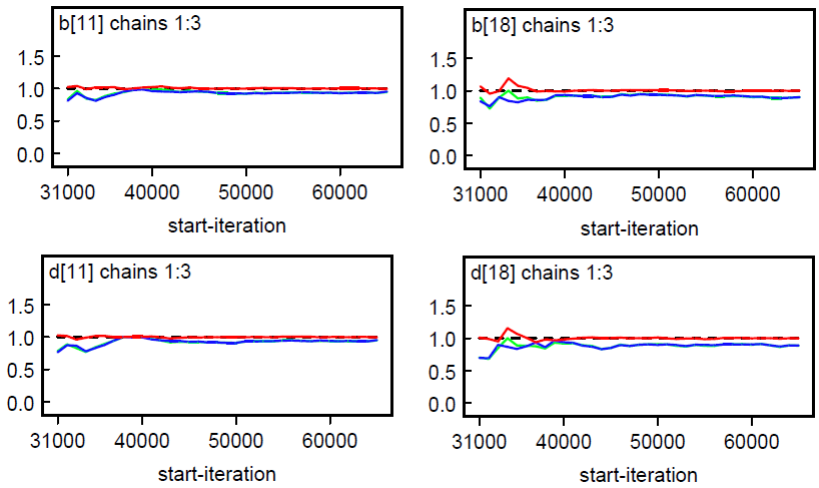


Figure 7 – Gelman Rubin convergence “Rasch" LPE for 36 items

Figure 6 outs some convergence problems for certain items (the green and blue curves should be stable), but, after removing item 28, the Gelman and Rubin statistic had improvements (Figure 7). For the other items, the Gelman-Rubin convergence statistics show that they have similar stability between both cases and that some items had convergence problems.

Results related with the Auto Correlation Function (not shown here) points out that the chains are highly auto correlated and one way to reduce the auto-correlation would be raising the thinning. With an independent chain, it could be possible to estimate the precision of the MCMC estimates. However, Link e Eaton (2012) studied the effects of thinning in MCMC chains using the Winbugs software. They claim thinning is not always necessary and that the thinned estimates are often not closer to the true value than the unthinned estimates. Additionally, because of the computational cost it would be very time consuming to raise the thinning to solve the auto-correlation problem.

These facts evidences a high computational cost and the need for a large sample of individual responses for the “Rasch" LPE model. This is also supported by the simulated results in the previous section.

Using the item parameters estimates as known values, the EAP method was implemented to estimate the latent traits for each case. The results were compared with the GS test score, which is the sum of correct responses. The correlation between the GS score and the EAP estimated was calculated and it is presented in Table 13.

Table 13 – Correlation between the GS score and θ estimate for each case

	Case 1	Case 2	Case 3
Correlation	0.9686	0.9857	0.9896
	Case 4	Case 5	Case 6
Correlation	0.9878	0.9892	0.9878

The GS score and EAP estimates are highly correlated in all cases, as seem in Table 13. Numerically, the correlation between the GS score and all “Rasch" LPE cases were slightly greater than the 2PL’s.

The kernel density distributions of GS score and θ are presented in Figure 8. The kernel density distributions is a non-parametric method to estimate the parameter density distribution, useful when their parametric distribution is unknown. This process requires a smoothing function (a non-negative function that integrates 1 and has mean 0) and positive parameter (h) also called bandwidth Rosenblatt (1956). In this article the standard normal distribution and h = 1 were chosen as the smoothing function and bandwidth parameter, respectively.

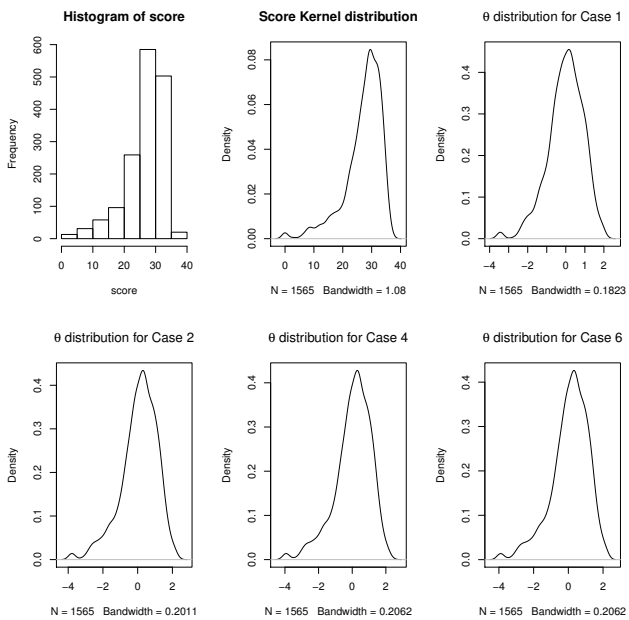


Figure 8 – Comparison of the individual sample score and distribution

The figure 8 and Table 13 show that even though “Rasch" LPE had issues of convergence, the latent trait estimates distribution are very similar to the 2PL and the total score.

The item fit analysis shown in this article does not correspond to all 6 cases mentioned above. Instead, the item fit was made for the MCMC estimates of Rasch, 2PL and “Rasch" LPE models using the responses of 36 items (item 28 was removed) of the GS test. The priors considered (if the parameter is present in the model) were: $Normal(0,1)$ for θ, $Normal(0,2)$ for b, $Log-Normal(0.1,\sqrt{0.5})$ for a and $Gamma(0.25,0.25)$ for λ. Notice that Case 6 was contemplated here.

The Rasch model was included in this analysis for comparison purpose, because both Rasch and “Rasch" LPE do not have the a parameter.

No item fit analysis were made for the MML method because of its slow convergence in the real data case.

For this fit analysis, 500 set of parameter values were selected from one of the 3 chains generated. These values were taken from every 20th sample after the 100000th chain position. Using these parameter samples, 500 replicates response data set were simulated. After that, the discrepancy measure $S-X^2$ and its ppp-values were calculated.

In Table 14, it is shown the estimated ppp-values for all items for the “Rasch" LPE model.

Table 14 – “Rasch" LPE ppp-values using $S-X^2$ as discrepancy

Item	ppp-value	Item	ppp-value	Item	ppp-value	Item	ppp-value
1	0.61	2	0.42	3	0.08	4	0.31
5	0.24	6	0.15	7	0.12	8	0.14
9	0.01	10	0.71	11	0.85	12	0.82
13	0.80	14	0.39	15	0.54	16	0.47
17	0.10	18	0.73	19	0.61	20	0.06
21	0.00	22	0.03	23	0.28	24	0.08
25	0.68	26	0.81	27	0.45	28	0.30
29	0.37	30	0.82	31	0.55	32	0.61
33	0.40	34	0.06	35	0.43	36	0.43

Notice in Table 27 that items 9, 21 and 22 have ppp-value below 0.05, indicating that the model has not a good fit for theses items. Item 20’s ppp-value is close to 0.05 and, in some instances, it is also included in item misfit subset.

The graphics of the observed score frequency (blue line), the mean (black line) and the credibility interval (gray area delimited by red lines) of the replicates score frequency, for the three items (9, 21 and 22) with ppp-value below 0.05 are displayed in Figure 9’s first row. In the second row, the graphics of three items with ppp-value above 0.05 were used for comparison purpose.

The graphics of items 9, 21 and 22, shown in Figure 9, have four, four and three points, respectively, distinguishably outside the CI with $\alpha = 95\%$. In items 15, 27 and 31 observed

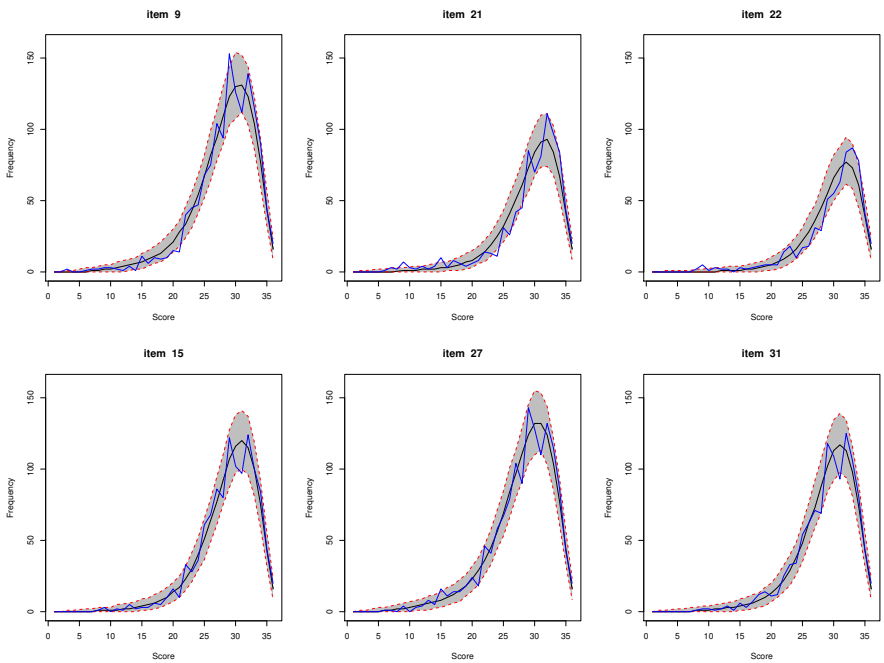


Figure 9 – Observed and replicated frequencies for “Rasch" LPE model

frequencies are inside or closer to the credibility interval with one, two and one point, respectively, clearly outside the interval.

Furthermore, items 9, 21 and 22 have patterns of score values that seems to be systematic, several consecutive points under or over the mean replicates frequency score (black line). Item 6 had this problem also. However, none of its systematic points were outside of the CI. In items 15, 27 and 31 no patterns were observed.

The graphics of other items (not shown here) also present points distinguishably out the CI (one to four points), most of them with just one, but they did not present the systematic pattern mentioned above.

The same item fit analysis were made for the Rasch and 2PL models and compared with “Rasch" LPE. For the Rasch model, the ppp-values are presented in Table 15 and the graphics of frequencies for the same items considered Figure 9 are displayed in Figure 10.

Table 15 shows that a subset of 10 items had ppp-value lower than 0.05, indicating that the Rasch model item fit was poor. This subset also contains the three poor fitted items from “Rasch" LPE (Items 9, 21 and 22).

In Figure 10, the item misfit is more perceivable for Rasch than the "Rasch" LPE. Item 21 had a much larger confidence interval for Rasch model than for the previous one, this fact also occurs to items 25, 28, 29, 32, 34 and 35.

Table 15 – Rasch ppp-values using $S - X^2$ as discrepancy

Item	ppp-value	Item	ppp-value	Item	ppp-value	Item	ppp-value
1	0.14	2	0.08	3	0.15	4	0.52
5	0.04	6	0.03	7	0.14	8	0.25
9	0.03	10	0.55	11	0.49	12	0.41
13	0.88	14	0.30	15	0.63	16	0.27
17	0.08	18	0.41	19	0.48	20	0.03
21	0.00	22	0.00	23	0.01	24	0.18
25	0.74	26	0.83	27	0.46	28	0.23
29	0.00	30	0.25	31	0.57	32	0.35
33	0.26	34	0.00	35	0.01	36	0.14

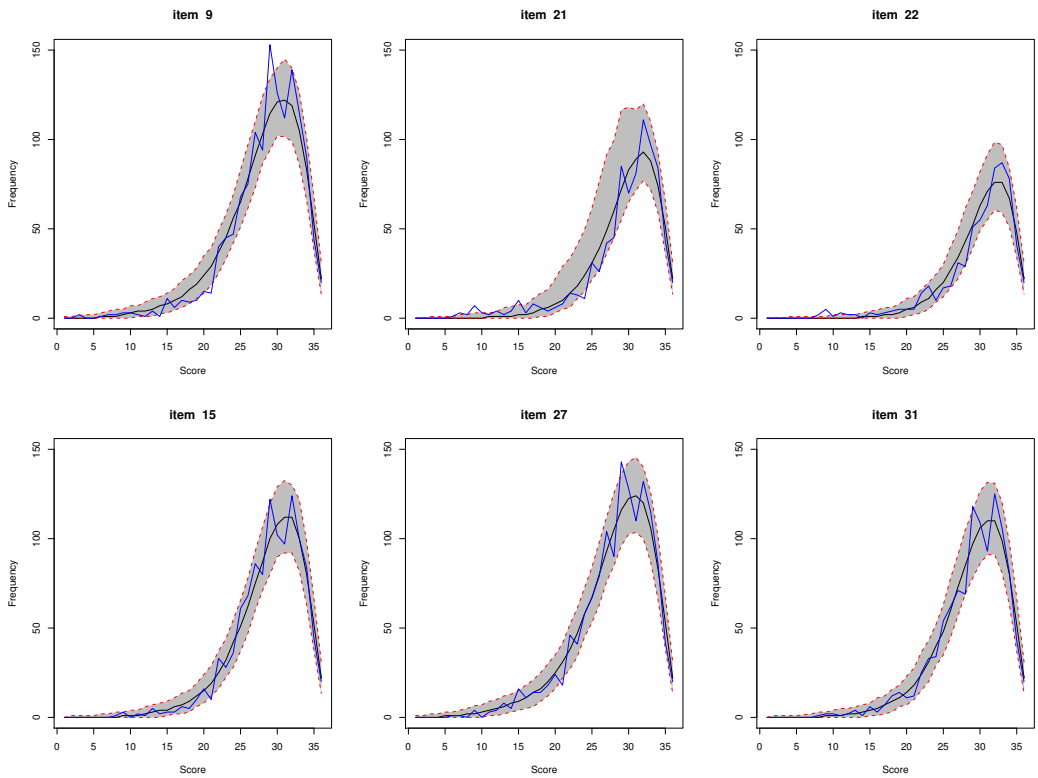


Figure 10 – Observed and replicated frequencies for Rasch model

Table 16 and Figure 11 presents the ppp-value and graphic of frequencies, respectively, for the 2PL model. Notice that the items in Figure 11 are the same from Figures 9 and 10.

In Table 16, differently from “Rasch" LPE, the ppp-values didn’t indicate a bad fit for item 22. It is also observed an improvement for items 17, 20, 24 and 28. Overall, the ppp-values points out that the 2PL model was just slightly better fit to the real data than “Rasch" LPE.

For the 2PL model (Figure 11), it is noteworthy that item 22 has a better fit than in “Rasch" LPE, with just one point outside the interval.

In Figure 9 and Figure 11, items 21, 22 and 31 had a better fit for the 2PL than “Rasch"

Table 16 – 2PL ppp-values using $S - X^2$ as discrepancy

Item	ppp-value	Item	ppp-value	Item	ppp-value	Item	ppp-value
1	0.90	2	0.59	3	0.14	4	0.46
5	0.22	6	0.21	7	0.18	8	0.14
9	0.00	10	0.71	11	0.94	12	0.93
13	0.91	14	0.44	15	0.50	16	0.81
17	0.74	18	0.75	19	0.54	20	0.65
21	0.00	22	0.07	23	0.56	24	0.25
25	0.72	26	0.91	27	0.37	28	0.61
29	0.43	30	0.84	31	0.58	32	0.72
33	0.51	34	0.15	35	0.68	36	0.63

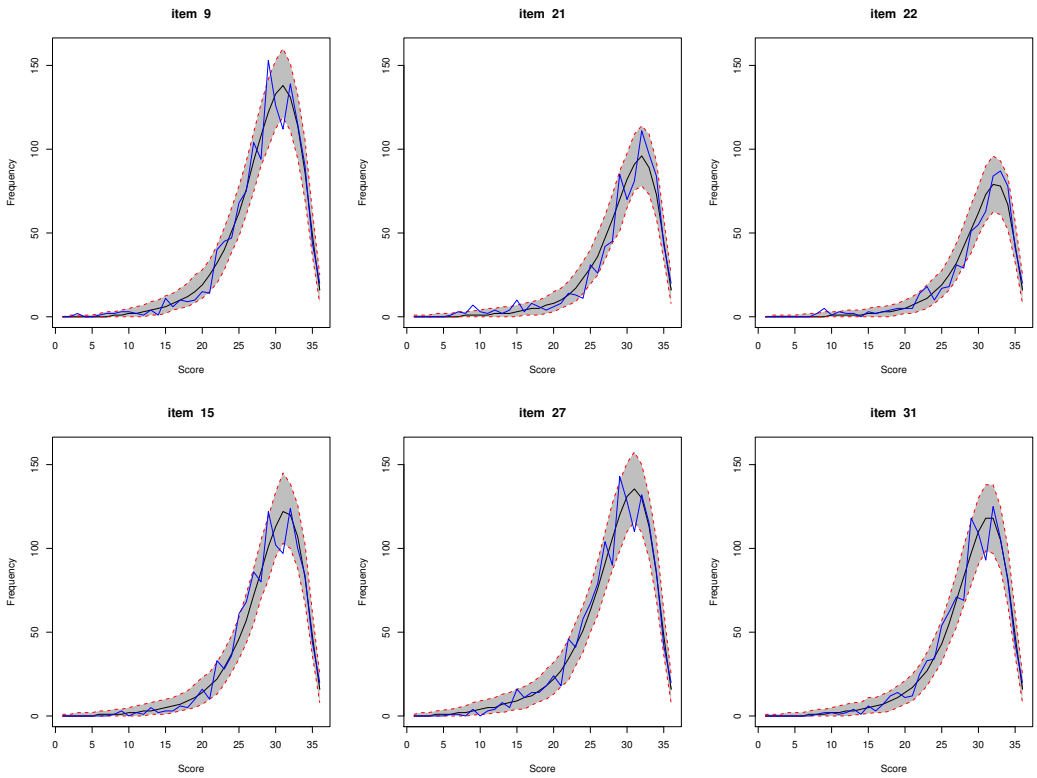


Figure 11 – Observed and replicated frequencies for '2PL model

LPE by having less points outside the CI and the points outside are closer to the limit bounds in the 2PL. However, “Rasch" LPE seems to have a slightly better fit for the items 9, 15 and 27.

It is noticeable, by comparing the graphic of estimated score frequencies between 2PL and “Rasch" LPE models, that all graphics for the 2PL model are dislocated slightly to the right in relation to the “Rasch" LPE model. This means that the items estimated under the 2PL model are easier than the “Rasch" LPE's.

In fact, “Rasch LPE" is consistently better than 2PL in the region around the score 30 (28 to 32), but worse outside.

The item fit analysis indicates that the 2PL model had a slightly better fit than the “Rasch" LPE and that the Rasch model had the worst fit. It is important to remind that the “Rasch" LPE model requires a greater individual sample size to estimates its parameters and that the “Rasch" LPE seemed to be more adequate in certain occasions. Lastly, the full LPE model was not considered because of difficulty in estimation.

Another criteria to compare model performance were made by using the information statistics. The E-AIC, E-BIC and DIC for the Rasch, 2PL and “Rasch LPE" models were calculated and the results are shown in Table 17:

Table 17 – Comparison of models

Model	EAIC	EBIC	DIC
Rasch	46802.48	46995.29	46741.16
2PL	45821.14	46206.74	45673.18
"Rasch" LPE	46329.15	46714.76	46216.18

Table 17 shows that for all 3 model comparison criteria, both 2PL and “Rasch" LPE had better performance than the Rasch model and the 2PL had the best performance considering the information statistics.

CHAPTER

7

MST RESULTS

In this chapter, it will be explained the module assembly process for the multistage adaptive test for one simulated and one real datasets.

The objective is to build a single panel for each information function proposed (F, KL and CEM) for the General Science (GS) test and a simulated dataset. The panel structure, module assembly process and results for each test are presented in the two sections below.

7.1 Simulated items

For the simulated item pool, 100 items were generated using the “Rasch" LPE model, the b parameters were withdrawn from a Normal(0,2) and λ from a Log-normal(-0.4,0.8) with expected value, variance, minimum and maximum of 1, 1.226, 0.057 and 4.255, respectively. The items parameters are show in Figures 18, 19 and 20.

For the multistage test simulation, the chosen panel structure was composed of 3 stages and 7 modules. The first stage had only the Routing module, the second and third stage contained 3 modules: the Easy, the Moderate and the Hard modules.

Each module had 10 items with no repetition. In this schema, 9 distinct tests with 30 items each can be formed for different ranges of individual performances.

To calculate F and KL informations, three fixed θ points were used: -1.5, 0 and 1.5, the most informative items were put in third stage’s Easy, Moderate and Hard module, respectively. This procedure was repeated once with the remaining items; the most the informative for each point composed the second stage’s Easy, Moderate and Hard module. Finally, for the Routing module, 5 points were used instead: -2,-1, 0, 1 and 2, and the 2 most informative for each point were selected to build it. However, the module information curve pointed out that the Routing module was on the easy side. So, one item for $\theta = -1$ and three items for $\theta = 0$ were chosen instead.

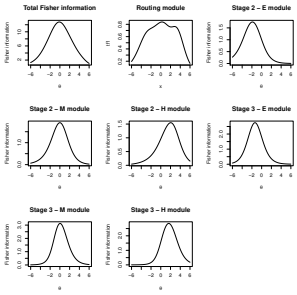


Figure 12 – Fisher module assembly method

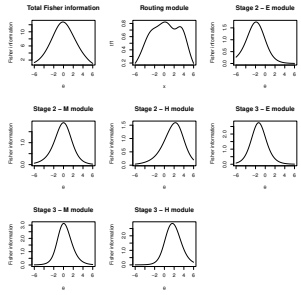


Figure 13 – KL module assembly method

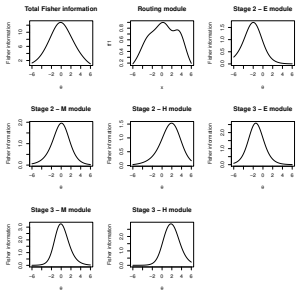


Figure 14 – CEM module assembly method

It is worthy note that the integration interval on KL information was $\pm 0.5, 1, 1.5$ for modules on stage three, two and one, respectively, because (reference of the KL suggested integration interval).

For CEM, instead of θ, the fixed values were attributed for μ and the selection were made in a similar manner as for F and KL. Remembering that in this method, items with minimum ECE are the ones to be selected.

In this assembly procedure, some of the 10 (or 2 for the Routing module) most informative items would appear in two or more modules at the same stage. In this case, one item would belong to the module in which it had most information (or least ECE for CEM), and the next most informative items would be selected for the other modules.

The Fisher information of the resultant modules are presented in Figure 12, 13 and 14. Because of the design specifications and the generated items, the following three characteristics can be observed for each information/entropy functions: later stages have more maximum information; the Easy, Moderate and Hard modules are optimal for different ranges of the ability scale and the Routing module information is relative homogeneous.

For the 9 possible simulated test, the Fisher information for each module assembly procedure is shown in Figures 15, 16 and 17. In this simulated scenario, the information curves across all paths are satisfactory, with each path having adequate amount of information for the aimed ability range.

In Figures 18, 19 and 20 are presented the bar plots illustrating the item parameters (red and blue bars) for each module assembled using Fisher, KL and CEM methods, respectively.

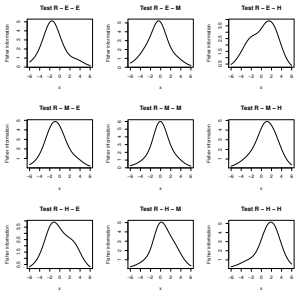


Figure 15 – Fisher module assembly method

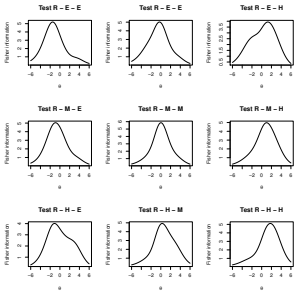


Figure 16 – KL module assembly method

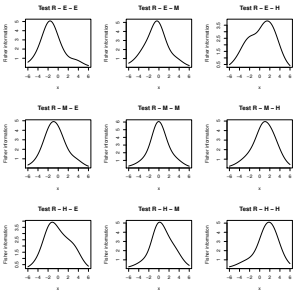


Figure 17 – CEM module assembly method

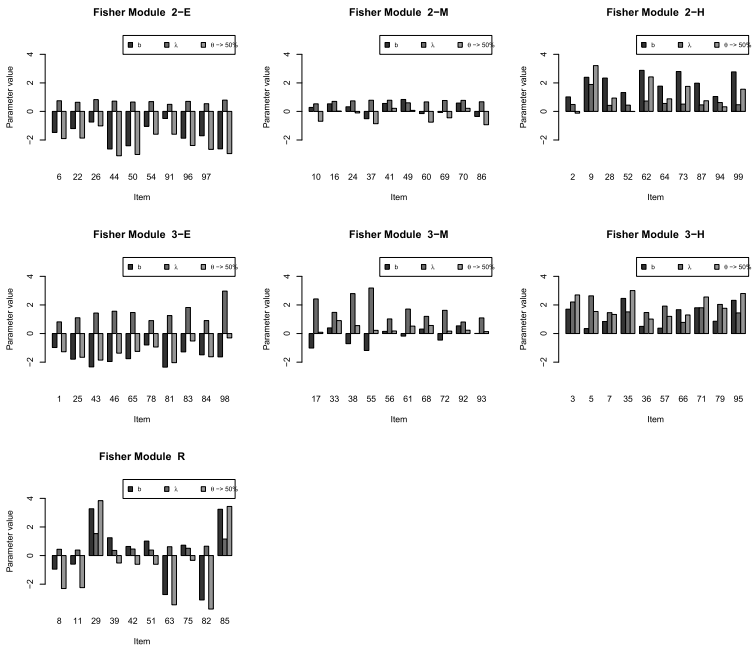


Figure 18 – Item parameter within each module

Additionally, the gray bars shows the θ value in which the probability of correct response for the item is 50%.

Because higher λ means greater information (or lesser ECE), items with higher λ values are found stage 3 for all methods. The Routing module has large range of b values because of the assembly design. As mentioned before, the information functions (or ECE), under the LPE model, aren't optimal for $\theta = 50\%$, therefore, it is not strange to have some items placed in

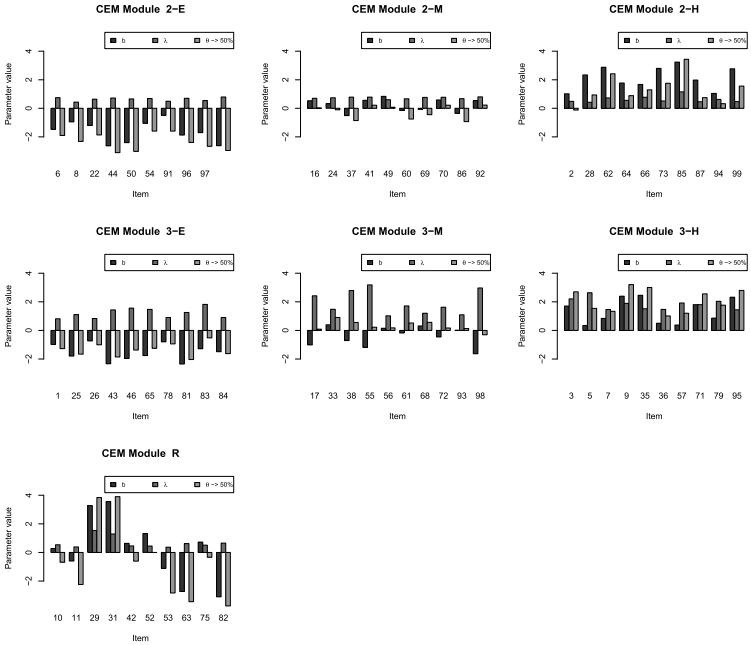


Figure 19 – Item parameter within each module

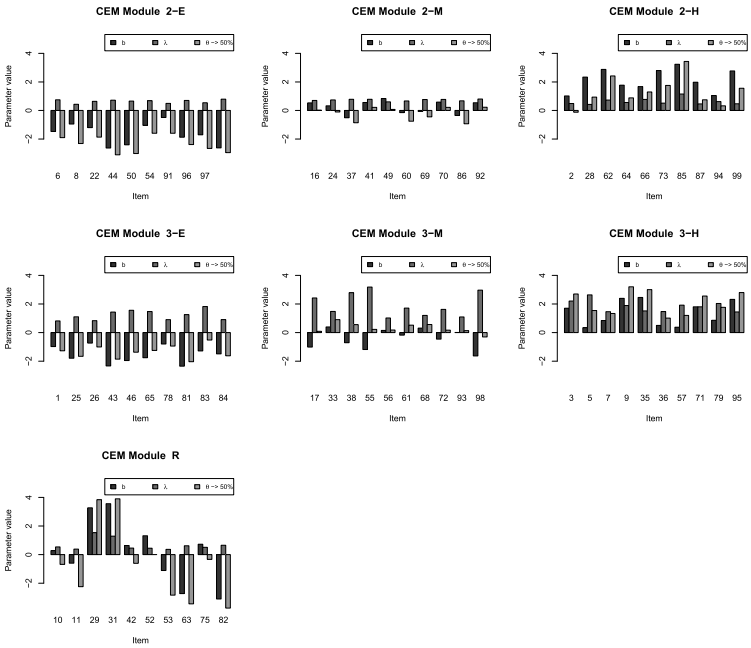


Figure 20 – Item parameter within each module

the Hard module even though their gray bar are below 0 (analogous for some items in the Easy module).

To compare the Rasch and 2PL with "Rasch" LPE, those were fitted to the response data generated by the "Rasch" LPE model, the estimates were obtained using the mirt package from R.

The priors for both Rasch and 2PL model were $a \sim N(-0.1, 4)$ (for the 2PL only), $b \sim Normal(0, 2)$, and $\theta \sim Normal(0, 1)$. A burn-in of 10000, 30000 iterations and thinning of 5 were used to obtain the estimates, Gelman and Rubin statistics indicates convergence for all chains and no auto-correlation, tendency nor seasonality were observed.

For the module assembly, the same procedure were used as for the Rasch “LPE" case. Stage 3 had the items with most information for the previously determined latent trait points, and the Routing module had information for large range of θ and centered close to 0.

In Table 27 are presented the list of items in each module for each assembly criteria for “Rasch" LPE, Rasch and 2PL models.

According to the Table 18, all module assembly methods produced the same results for the Rasch model. This was expected since the information functions and CEM values are determined only by the difference between θ and b parameters. For the 2PL and “Rasch" LPE, all 3 panels formed are slightly different from each other.

In respect to the differences between the models, the Rasch panels are the most different from the others, even though they had some items (in general, less than 40%) in common. Although the "Rasch" LPE and 2PL produced very similar items for stage 3 modules, they have 7-9 out of 10 items in common. the similarity diminishes for stage 2 and Routing modules, 4-7 out of 10. However, some of those "missing items" can appear in the same stage but different difficult levels (item 33, 83, etc...) or appears in same difficult levels but in different stage (item 6, 92, etc...).

With the modules and panels built, the next step is to estimate individual abilities.

For the latent trait estimation, the EAP method was implemented. In this process, the standard normal distribution were assumed as θ's prior and the generated responses were used as the data. The EAP algorithm was applied for test with all 100 generated items and the multistage test (30 items).

To determine which module should be applied next, three different methods were initially considered: 2 of them were based on specific cut points and the third one selected the module which the most Fisher information for the θ estimate. However, in our study, the latter had a considerable worse performance than the cut points for the “Rasch" LPE and it will not be considered for other cases in this dissertation.

The 2 sets of cut points which their respective Bias and RMSE results for each model is presented in Table 19, in this table, the Bias and RMSE for 100 items is also available.

In 19, Easy cut points means that, after completing a module, if an individual had his θ estimate below than the Easy cut point value, he would be presented with the Easy module. On opposite side, if an individual obtained a θ estimate above the Hard cut point value, he would be presented with the Hard module. If none of the previous situations was applied, the Moderate module was administered instead.

The θ estimates for the 2PL and the complete item had a small significant bias. However, in the multistage tests, only the Rasch model estimates were biased. In relation to the assembly methods, CEM had the lowest RMSE for the “Rasch" LPE model, but the highest for the 2PL model. The opposite effect can be observed for Fisher information.

Another focus of this dissertation is to analyze the probability of correct response, this is important to visualize which assembly method, under each of the presented models, administered the most appropriate set of items for the examinees.

For that purpose, the average probability of correct response for each stage was calculated

as follow: for the Routing stage, the probabilities of correct response for the Routing modules were calculated based on the θ estimates using the response from the Routing module; for Stage 2, the probabilities of correct response for the selected modules presented in the Stage 2 were calculated based on the θ estimate using the responses feom the two first stages; finally, for Stage 3, the probability of correct response for the selected module presented in the Stage 3 was calculated based on the θ estimate using all stages.

The resultant average probabilities for each model, module assembly method and cut points are presented in Table 20.

Since every module is the same for the Rasch model, there are no average probability differences between any assembly method within each cut points. Additionally, the 2PL model differences are marginal. However, for the “Rasch" LPE, the average probabilities had significantly distinctions (based on t-test) in stage 3 (the final and most informative modules) as seem in Table 20. For both cut points, The CEM presented Stage 3 modules that were more adequate to the individual‘s estimated abilities than the other two methods.

7.2 Guessing and Accidental Mistakes

Two other studies were made using the data from the previous section to analyze the 2PL’s and “Rasch" LPE’s (the Rasch model wasn’t analyzed in this scenario because it had a poor performance on the previous section) latent trait estimates in relation to guessing and accidental mistakes. In the first study, the wrong responses were altered to simulate guessing. Randomly, 20% of all wrong responses (0) were changed to right answers (1). Secondly, 10% of all right responses were considered wrong to simulate accidental mistakes.

For both studies, the same MST structure from the previous section was adopted. A single panel was used and it was composed of 3 stages and 7 modules. The first stage had only the Routing module, the second and third stage contained 3 modules: the Easy, the Moderate and the Hard modules.

The F and KL informations and CEM were used to select items. However, some constraints were included in this case: For the Hard modules, only the 30 items with highest a (for 2PL) or λ (for LPE) and $\theta_{50\%} > 0$ (defined in Chapter 2) were considered as candidate. For the Easy modules, only the 30 items with lowest a or λ and $\theta_{50\%} < 0$ were considered as candidate. The remaining items were candidates for the Moderate and Routing modules.

In this setup, the most informative items were put in the Hard modules. This ordering can be justified for tests where the interest is to select high proficient individuals (classification tests).

With the structure of the test built, the EAP method was implemented to estimate the latent traits using the modified response patterns. Only the ± 1 cut off points were considered

because it had the best performance in the previous section.

The results (Bias and RMSE) from the first study, that considered the scenario of individuals making 20% of correct guesses on items that they didn't know how to solve, are presented on Table 21.

Table 21 presents great negative bias (true - estimated values) for all cases. This was expected because some of the wrong responses were altered to right.

For the complete test, the 2PL had slightly less bias and RMSE. For the MST's θ estimates for the "Rasch" LPE for Fisher's and KL's test were farther from the true value than 2PL's. However, for the CEM's, the two models had similar results. The CEM assembly method had the best performance among the assembly methods.

The resultant average probabilities for each model, module assembly method and cut points are presented in Table 22.

Table 22 shows that KL based module assembly produced average probabilities closer to 50% on stage 3 for both models. For stage 2, Fisher's had the better performance. In stage one, the differences between assembly methods are minor. On the Routing stage CEM was the closest for the 2PL model.

The second study was realized to evaluate the effects of accidental mistakes for the MST built in this section. The results are presented in Table 23, in this table, the Bias and RMSE for 100 items is also available.

For the complete test, the 2PL was slightly less bias but similar RMSE. For the MST's θ estimates for the "Rasch" LPE for Fisher's and KL's test were more distant from the true value than 2PL's. However, for the CEM's, the opposite effect can be observed. The Fisher assembly method had the best performance among the assembly methods.

The resultant average probabilities for each model, module assembly method and cut points are presented in Table 24.

As in the previous study, Table 24 shows that the KL based module assembly produced average probabilities closer to 50% on stage 3 for both models. For stage 2, Fisher had slightly better performance. On the Routing stage CEM was the closest for the 2PL model.

7.3 Comparison of the LPE and 2PL models

A simulation study comparing two multistage tests using the LPE with a parameter not fixed at 1 and 2PL models was made.

The previous scenario was considered (100 items and 5000 individuals). The data was generated using the LPE model with b and λ withdrawn from the same distributions of the previous subsection. The a parameter was generated using a $log-normal(-0.1, 0.25)$

distribution with mean 1.025 and variance 0.287, the maximum value was 3.89 and minimum value was 0.28.

The Rasch and 2PL model parameters were estimated using a Bayesian approach and the MCMC method. The same previous priors for these models were considered. The chain size, burn-in and thinning were maintained.

The multistage structure, module assembly and cut points were determined equally as in the previous subsection.

The EAP estimation were made for the full length test (100 items) and for the multistage tests with module assembly based on Fisher, KL and CEM functions. The results are shown in Table 25

In Table 25, the 100 items θ estimates for the 2PL and the LPE models have very similar efficiency, even though the responses were generated under the LPE model. However, in a multistage context, the LPE had lower RMSE than 2PL. It is also worth to notice that, for the LPE, some slightly significative bias, but that was not the case for the CEM bias. Rasch model had a worse performance in overall.

The average probabilities of correct response was also calculated in this simulation, the results are presented in Table 26.

Table 26 shows that the KL based module assembly produced average probabilities slightly closer to 50% for the $\pm$ 0.75 cutting point than Fisher's and CEM's under the "Rasch" LPE model and for the third stage of 2PL for both cutting points. For all other cases, performance of the 3 module assembly criteria were very similar.

7.4 Real data application

The General Science (GS) test is composed by 37 items adapted from the SLiM Instrument Rundgren *et al.* (2012) about the general science knowledge with a reliability (Cronbach's α) of 0.89. However, only 36 items were used (item 28 removed from the pool due to its negative total correlation) to build the multistage test.

To assemble the multistage test, first, it was defined that the panel structure would contains 3 modules and 2 stage. The first stage would have only the Routing module and the second stage would contain 2 modules: the Easy and the Hard modules.

Each module had 12 items, no item were used more than once and every item of the pool were used. The 2 possible multistage tests are composed of 24 items each.

To calculate Fisher and KL informations, two fixed θ points were specified: -1.5 and 1.5, the most informative items for $\theta = -1.5$ and 1.5 were put in the Easy and Hard module, respectively. The remaining items composed the Routing module.

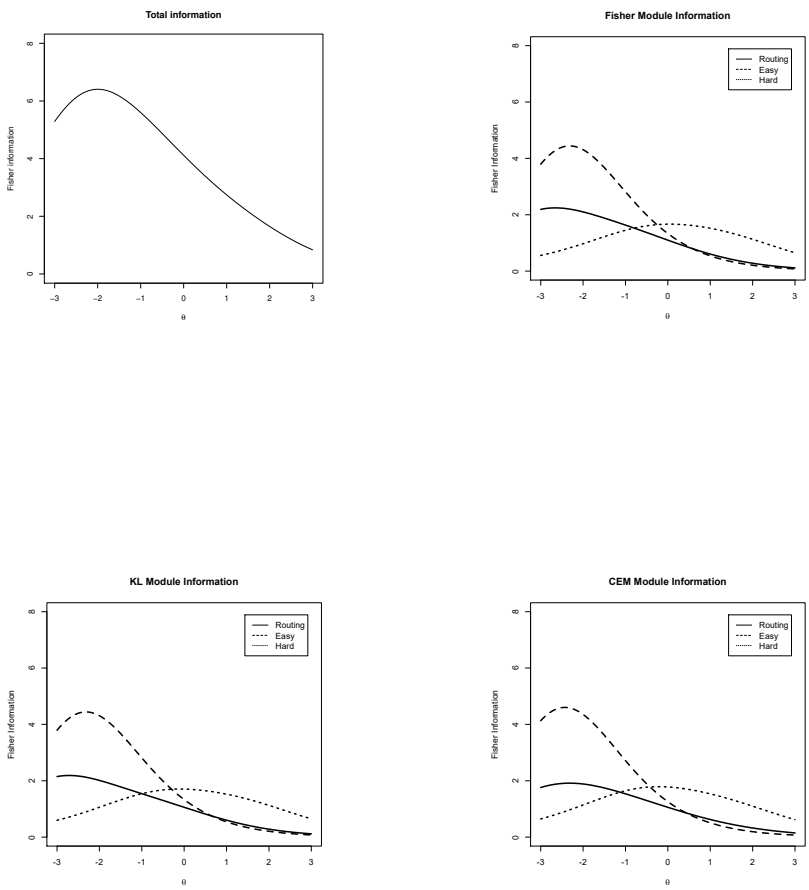


Figure 21 – Fisher information for the complete GS test and each module assembly criteria

It is worth noting that the integration interval on KL information was ± 1. However, if the interval were $\pm$ 0.5 or lower, the two panels generated for each information would be the same. This is a reminder that for small integration intervals, F and KL tends to be equivalent.

For CEM, θ is assumed to have a Normal prior distribution with mean $= \mu$ and variance $= 1$. Since ECE doesn't depends on θ directly, the -1.5 and 1.5 values were attribute for μ instead. In a similar manner, items with the least ECE values for $\mu = -1.5 and 1.5$ were put in the Easy and Hard module, respectively, and the remaining items in the Routing module.

In Figure 21, for each assembly procedure, is presented the Fisher information (for comparison purpose) for the Rounting, Easy and Hard modules. The Total information reveals the GS test as being an easy test, with low capability to differentiate individuals with high general science knowledge very well.

Note in figure 21 that the CEM's Easy module has lower Fisher information for high

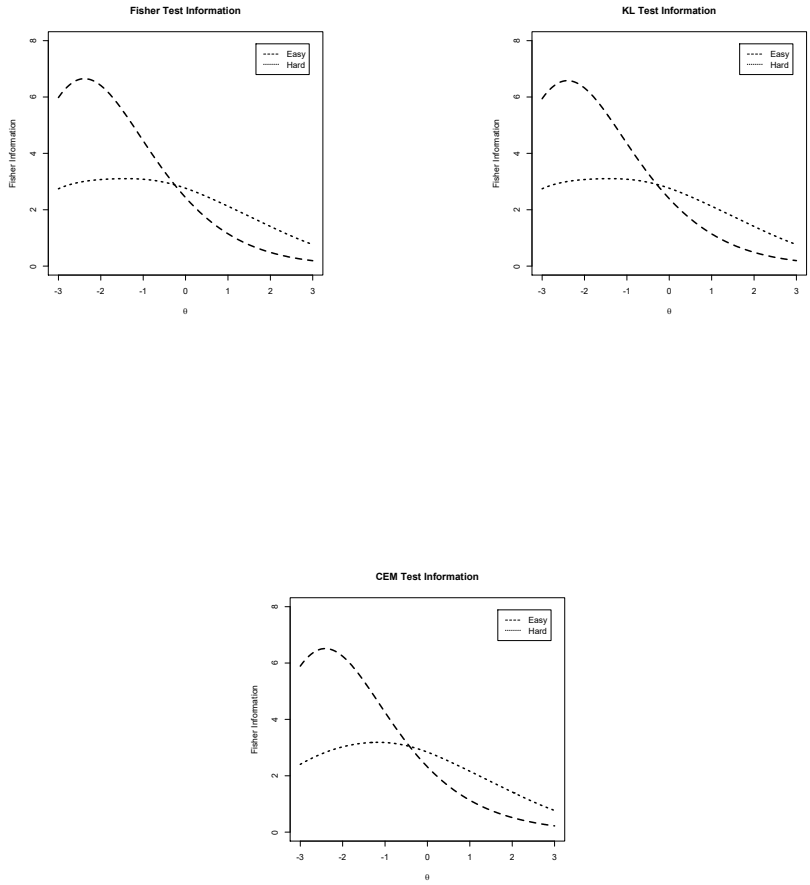


Figure 22 – Fisher information for both paths of the GS multistage test for each module assembly criteria

θ values than the other Easy modules. For the GS items, the CEM produced the most singular panel.

Because the GS test is not very informative for higher values of θ, its multistage test had similar characteristic, as evidenced in Figure 22. However, there are a clear distinction between the Easy and Hard path of the multistage test, the hard path provides considerably more information for higher abilities. Based in the graphics, a cut point of $\theta = -0.5$ in the Routing module can be established to route the examinees to the easy or hard module, additionally, the mid point of the ability scale, $\theta = 0$ will also be used as cut point.

The Figure 23 presents bar plots illustrating the item parameter values (blue and red bars) and for each module, additionally, the gray bars shows the θ value in which the probability of correct response for the item is 50%.

In “Rasch" LPE, λ value is the only parameter that affects the maximum Fisher informa-

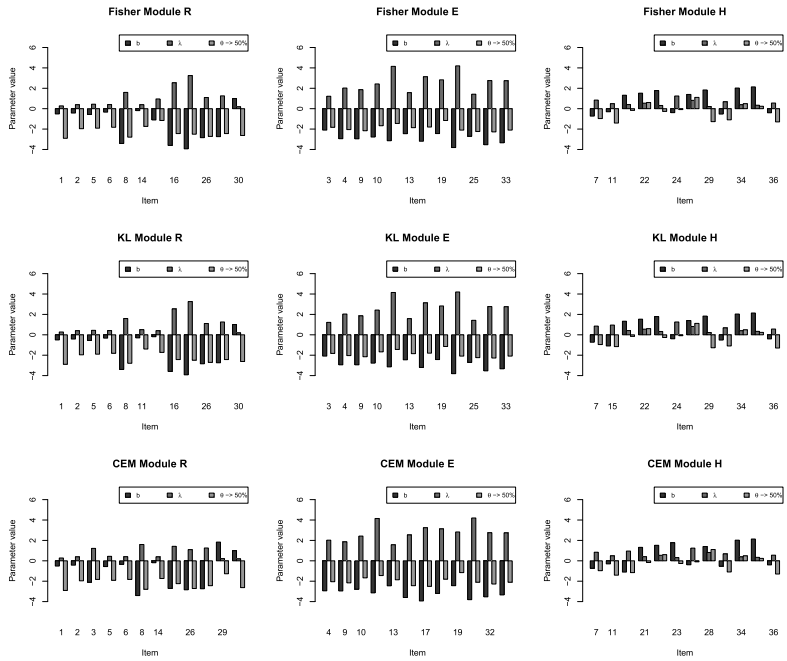


Figure 23 – Item parameter within each module

tion value. For the GS test, items with high difficulty had low values of λ, as shown in Figure 23, moreover, around 50% of items have $\lambda < 0.9$ and only 2 items have $\lambda > 1.1$ for the Hard mode, while all items on Easy module have $\lambda > 1.3$. This explain why the total information for the hard module is lower than the Easy module.

To compare the Rasch and 2PL with "Rasch" LPE, those were fitted to the response data for the GS test also, the estimates were obtained using the WinBUGS software through a Bayesian MCMC approach.

The priors for both Rasch and 2PL model were $a \sim N(-0.1, 4)$ (for the 2PL only), $b \sim Normal(0, 2)$, and $\theta \sim Normal(0, 1)$. A burn-in of 20000, 50000 iterations with thin of 5 where used to obtain the estimates, Gelman and Rubin statistics indicates convergence for all chains and no auto-correlation, tendency nor seasonality were observed.

For the module assembly, the same procedure were used as for the Rasch “LPE" case. Stage 2 have the items with most information for the previously determined latent trait points and the remaining items were assign to the Routing module.

In Table 27 are presented the list of items in each module for each assembly criteria for “Rasch" LPE, Rasch and 2PL models.

According to the Table 27, all module assembly methods produced the same results for the Rasch model for the same reason mentioned in the simulated case. For the 2PL, KL and CEM produced the same panels and the only difference between them and fisher was in the Routing module. Lastly, in “Rasch" LPE, all 3 panels formed are slightly different from each other.

In respect of the panel differences between the models, the Rasch’s are the most different from the others with some items in common, around 1/3, 1/4 and 2/3 items in common for the Routing, Easy and Hard module, respectively. The "Rasch" LPE’s and 2PL’s produced very similar items for all modules, differences of 1 to 3 items for each module.

The EAP method was implemented for the latent trait estimation. In this process, the standard normal distribution were assumed as θ’s prior and the generated responses were used as the data. The EAP algorithm was applied for test with all 36 generated items and the multistage test (24 items).

To determine which module should be applied next, two cut points were considered: $\theta = 0$ and $\theta = -0.5$ which means that If an individual had θ estimate below those values, they were presented with the Easy module, otherwise, the Hard one was administered.

The 2 sets of cut points which their respective Bias and RMSE results for each model is presented in Table 19, in this table, the Bias and RMSE for 100 items is also available.

In table 28 shows that the “Rasch LPE" multistage test is the most different from its full test, meaning, that for this case, more items in the multistage test would be required for this model to have as close results to the full length test as the 2PL model.

The correlation between θ estimates for each model using all items is very high, the greatest distance of the estimates is between the 2PL and “Rasch" LPE, but still very close.

In the probability of correct response perspective, the average probabilities were calculated in the same fashion as the simulated case.

The resultant average probabilities for each model, module assembly method and cut points are presented in Table 29.

In Table 29, the cut point of -0.5 have more appropriate average probabilities for all models, and because in Table 28, there are no significative difference between the two cut points, $\theta = -0.5$ produced the best results.

Between each module assembly method, although the CEM had slightly lower probability average, no significative differences were found. Notice that not only “Rasch" LPE, but also the Rasch and 2PL had high average of probability of correct response due to the easiness of the test.

Table 18 – List of items in each module for each model and assembly method

	Rasch	2PL			"Rasch LPE"		
	All	Fisher	KL	CEM	Fisher	KL	CEM
Module R	3	2	2	2	8	8	10
	9	13	13	13	11	11	11
	10	31	31	28	29	29	29
	26	39	39	39	39	31	31
	35	51	51	51	42	39	42
	42	52	52	52	51	42	52
	51	74	74	74	63	51	53
	80	75	75	75	75	63	63
	81	96	82	82	82	75	75
	85	99	99	99	85	82	82
Module 2-E	1	8	8	8	6	6	6
	6	11	11	11	22	22	8
	8	22	22	22	26	26	22
	11	26	26	26	44	44	44
	15	37	37	37	50	50	50
	43	43	43	43	54	54	54
	48	53	53	53	91	91	91
	76	54	54	54	96	96	96
	77	86	86	86	97	97	97
	90	91	91	91	100	100	100
Module 2-M	39	10	10	10	10	10	16
	55	16	16	16	16	16	24
	69	24	24	24	24	24	37
	70	41	41	41	37	37	41
	72	42	42	42	41	41	49
	75	60	60	60	49	49	60
	83	69	69	69	60	60	69
	92	70	70	70	69	69	70
	94	92	92	92	70	70	86
	98	94	94	94	86	86	92
Module 2-H	19	9	9	9	2	2	2
	28	28	28	29	9	9	28
	38	29	29	31	28	28	62
	61	49	49	49	52	52	64
	64	62	62	62	62	62	66
	68	64	64	64	64	64	73
	71	66	66	66	73	73	85
	79	73	73	73	87	85	87
	87	85	85	85	94	87	94
	95	87	87	87	99	94	99
Module 3-E	4	1	1	1	1	1	1
	22	6	6	6	25	25	25
	25	25	25	25	43	43	26
	30	46	46	46	46	46	43
	46	65	65	65	65	65	46
	54	78	78	78	78	78	65
	59	81	81	81	81	81	78
	65	84	84	84	83	83	81
	84	97	97	97	84	84	83
	91	100	100	100	98	98	84
Module 3-M	2	17	17	17	17	17	17
	13	38	38	38	33	33	33
	16	55	55	55	38	38	38
	17	56	56	56	55	55	55
	24	61	61	61	56	56	56
	41	68	68	68	61	61	61
	49	72	72	72	68	68	68
	52	83	83	83	72	72	72
	56	93	93	93	92	92	93
	93	98	98	98	93	93	98
Module 3-H	5	3	3	3	3	3	3
	7	5	5	5	5	5	5
	33	7	7	7	7	7	7
	36	33	33	33	35	35	9
	57	35	35	35	36	36	35
	62	36	36	36	57	57	36
	66	57	57	57	66	66	57
	73	71	71	71	71	71	71
	74	79	79	79	79	79	79
	99	95	95	95	95	95	95

Table 19 – Full and MST's Bias and RSME θ estimations for simulated data

			Fisher		KL		CEM	
Model	Easy cut points	Hard cut points	θ Bias	θ RMSE	θ Bias	θ RMSE	θ Bias	θ RMSE
Rasch	-0.75	0.75	0.0157	0.4340	0.0157	0.4340	0.0157	0.4340
	-1	1	0.0138	0.4294	0.0138	0.4294	0.0138	0.4294
2PL	-0.75	0.75	0.0082	0.4068	0.0086	0.4095	0.0081	0.4132
	-1	1	0.0090	0.4043	0.0089	0.4067	0.0070	0.4073
"Rasch" LPE	-0.75	0.75	0.0082	0.4124	0.0111	0.4118	0.0072	0.4088
	-1	1	0.0038	0.4149	0.0079	0.4130	0.0076	0.4073
Rasch	100 items		Bias = 0.0072 ($=0$)			RMSE = 0.3044		
2PL	100 items		Bias = 0.0080 ($\neq 0$)			RMSE = 0.2849		
"Rasch" LPE	100 items		Bias = 0.0025 ($=0$)			RMSE = 0.2811		

Table 20 – Average probabilities for each model, assembly criteria and cut points for simulated data

		cut points $\pm$ 0.75			cut points $\pm$ 1		
		Fisher	KL	CEM	Fisher	KL	CEM
Routing stage	Rasch		0.4829			0.4829	
	2PL	0.5045	0.5045	0.5045	0.5045	0.5045	0.5045
	"Rasch" LPE	0.4777	0.4782	0.4772	0.4777	0.4782	0.4772
Stage 2	Rasch		0.6305			0.6682	
	2PL	0.6535	0.6539	0.6537	0.6741	0.6748	0.6786
	"Rasch" LPE	0.6443	0.6466	0.6497	0.6752	0.6782	0.6768
Stage 3	Rasch		0.4902			0.4896	
	2PL	0.4704	0.4704	0.4715	0.4681	0.4658	0.4658
	"Rasch" LPE	0.4079	0.4057	0.4191	0.4108	0.4116	0.4256

Table 21 – Complete test and MST θ estimates, Bias' and RSMEs for the Rasch, 2PL and "Rasch" LPE guessing simulation.

	Cut off points		Fisher		KL		CEM	
Model	Easy module $\theta <$	Hard module $\theta >$	θ Bias	θ RMSE	θ Bias	θ RMSE	θ Bias	θ RMSE
2PL	-1	1	-0.4251	0.4585	-0.4005	0.4665	-0.3825	0.4526
"Rasch" LPE	-1	1	-0.4462	0.4829	-0.4473	0.5150	-0.3884	0.4532
2PL	100 items		Bias = -0.6902			RMSE = 0.6505		
"Rasch" LPE	100 items		Bias = -0.7111			RMSE = 0.6793		

Table 22 – Average probabilities for each model, assembly criteria and cut points for guessing simulation

		Guessing		
		Fisher	KL	CEM
Routing stage	2PL	0.5188	0.5216	0.5096
	"Rasch" LPE	0.5132	0.5252	0.5148
Stage 2	2PL	0.6898	0.7174	0.7214
	"Rasch" LPE	0.6848	0.7005	0.7062
Stage 3	2PL	0.6533	0.5518	0.5979
	"Rasch" LPE	0.5787	0.5395	0.6057

Table 23 – Complete test and MST θ estimates, Bias' and RSMEs for the Rasch, 2PL and "Rasch" LPE for mistake simulation.

	Cut off points		Fisher		KL		CEM	
Model	Easy module $\theta <$	Hard module $\theta >$	θ Bias	θ RMSE	θ Bias	θ RMSE	θ Bias	θ RMSE
2PL	-1	1	0.3251	0.3815	0.4204	0.5196	0.3824	0.4652
"Rasch" LPE	-1	1	0.2947	0.3490	0.3827	0.4644	0.4330	0.5246
2PL	100 items		Bias = 0.4014			RMSE = 0.2859		
"Rasch" LPE	100 items		Bias = 0.4112			RMSE = 0.2873		

Table 24 – Average probabilities for each model, assembly criteria and cut points for mistake simulation

		Mistake		
		Fisher	KL	CEM
Routing stage	2PL	0.4446	0.4464	0.4378
	"Rasch" LPE	0.4502	0.4540	0.4496
Stage 2	2PL	0.6449	0.6558	0.6619
	"Rasch" LPE	0.6443	0.6560	0.6536
Stage 3	2PL	0.6277	0.5040	0.5341
	"Rasch" LPE	0.5555	0.5157	0.5652

Table 25 – Complete test and MST θ estimates, Bias' and RSMEs for the 2PL and LPE comparison study.

	Cut off points		Fisher		KL		CEM	
Model	Easy module $\theta <$	Hard module $\theta >$	θ Bias	θ RMSE	θ Bias	θ RMSE	θ Bias	θ RMSE
Rasch	-0.75	0.75	0.0107*	0.5043	0.0107*	0.5043	0.0107*	0.5043
	-1	1	0.0094	0.4888	0.0094	0.4888	0.0094	0.4888
2PL	-0.75	0.75	-0.0040	0.4881	-0.0010	0.4974	-0.0006	0.4327
	-1	1	-0.0010	0.4325	0.0004	0.4333	-0.0003	0.4338
LPE	-0.75	0.75	0.0080	0.3814	0.0105*	0.3800	0.0063	0.3791
	-1	1	0.0112*	0.3761	0.0108*	0.3727	0.0078	0.3774
Rasch	100 items		Bias = 0.0012				RMSE = 0.3554	
2PL	100 items		Bias = -0.0005				RMSE = 0.2786	
LPE	100 items		Bias = 0.0041				RMSE = 0.2746	

* Hypothesis of Bias = 0 rejected.

Table 26 – Average probabilities for each model, assembly criteria and cut points for 2PL and LPE comparison study

		cut points ± 0.75			cut points ± 1		
		Fisher	KL	CEM	Fisher	KL	CEM
Routing stage	Rasch		0.5093			0.5093	
	2PL	0.5240	0.5241	0.5239	0.5240	0.5241	0.5239
	LPE	0.5346	0.5353	0.5356	0.5346	0.5353	0.5356
Stage 2	Rasch		0.7414			0.7513	
	2PL	0.7698	0.7710	0.7601	0.7690	0.7688	0.7683
	LPE	0.7796	0.7753	0.7772	0.7837	0.7811	0.7822
Stage 3	Rasch		0.3595			0.3218	
	2PL	0.3112	0.3307	0.2963	0.2424	0.2569	0.2383
	LPE	0.4476	0.4517	0.4476	0.4320	0.4320	0.4322

Table 27 – List of items in each module for each model and assembly method using GS test data

	Rasch	2PL			"Rasch LPE"		
	All	Fisher	KL	CEM	Fisher	KL	CEM
Moodule R	1	1	1	1	1	1	1
	4	2	2	2	2	2	2
	8	3	3	3	5	5	3
	9	5	5	5	6	6	5
	16	6	6	6	8	8	6
	17	13	11	11	14	11	8
	20	14	13	13	15	14	14
	25	19	14	14	16	16	25
	26	26	26	26	17	17	26
	27	27	27	27	26	26	27
	32	29	29	29	27	27	29
	33	30	30	30	30	30	30
Module E	2	4	4	4	3	3	4
	3	8	8	8	4	4	9
	5	9	9	9	9	9	10
	6	10	10	10	10	10	12
	10	12	12	12	12	12	13
	11	16	16	16	13	13	16
	12	17	17	17	18	18	17
	14	18	18	18	19	19	18
	15	20	20	20	20	20	19
	19	25	25	25	25	25	20
	30	32	32	32	32	32	32
	36	33	33	33	33	33	33
Module H	7	7	7	7	7	7	7
	13	11	15	15	11	15	11
	18	15	19	19	21	21	15
	21	21	21	21	22	22	21
	22	22	22	22	23	23	22
	23	23	23	23	24	24	23
	24	24	24	24	28	28	24
	28	28	28	28	29	29	28
	29	31	31	31	31	31	31
	31	34	34	34	34	34	34
	34	35	35	35	35	35	35
	35	36	36	36	36	36	36

Table 28 – GS test results of θ estimations: Bias and RSME

		Fisher		KL		CEM	
Model	Cut points	θ Bias	θ RMSE	θ Bias	θ RMSE	θ Bias	θ RMSE
Rasch	-0.5	-0.0033	0.4229	-0.0033	0.4229	-0.0033	0.4229
	0	-0.0013	0.4231	-0.0013	0.4231	-0.0013	0.4231
2PL	-0.5	0.0013	0.3762	0.0006	0.3858	0.0006	0.3858
	0	-0.0004	0.3748	-0.0009	0.3830	-0.0009	0.3830
“Rasch" LPE	-0.5	0.0426	0.4386	0.0445	0.4401	0.0450	0.4482
	0	0.0384	0.4389	0.0356	0.4384	0.0434	0.4444
Rasch, 2PL	36 items	Cor = 0.9965			Distance = 0.1614		
2PL, “R" LPE	36 items	Cor = 0.9889			Distance = 0.2036		
“R" LPE, Rasch	36 items	Cor = 0.9928			Distance = 0.1283		

Table 29 – Average probabilities for each model, assembly criteria and cut points using GS test data

		cut point = -0.5			cut point = 0		
		Fisher	KL	CEM	Fisher	KL	CEM
Routing stage	Rasch		0.8318			0.8318	
	2PL	0.7919	0.7921	0.7921	0.7919	0.7921	0.7921
	“Rasch" LPE	0.8272	0.8272	0.8277	0.8272	0.8272	0.8277
Stage 2	Rasch		0.7661			0.7777	
	2PL	0.7259	0.7266	0.7266	0.7501	0.7517	0.7517
	“Rasch" LPE	0.7397	0.7397	0.7394	0.7738	0.7725	0.7693

CHAPTER

8

CONCLUSION AND DISCUSSION

In this dissertation, we implemented and compared the LPE model with the most popular IRT models for dichotomous items responses. Some of these studies used a particular case of the LPE model, the “Rasch" LPE. The comparisons of the models were also explored in the MSTs scenarios. Additionaly, Fisher and KL information and CEM based module assembly methods were analyzed.

The data considered were three simulated and a real data set application, the General Science test from the Tetralogue project.

The LPE model ability to reward more the right answer for more difficulty items set it apart from the classic 2PL and 3PL models. This translates in the possibility of having items in the test, under the same model, that penalizes random guessing and items that help candidates to recover from early mistakes in easier items as seeing in Table 4, 18 and 27.

However, there are some practical issues to the model that makes the parameter estimation difficult and computational expensive. Simplifying the LPE by fixing the a parameter helps with the estimation problems, but removing it from the model, makes λ the only parameter that changes the inclination of the ICC, meaning that inclination and asymmetry are tied in the “Rasch" LPE model. Considering a large pool of items, these asymmetric models (LPE and “Rasch" LPE) would fit better in tests that have fewer guessing rates and greater stress related mistakes.

Two estimation methods were implemented for some of the analysis: a Bayesian MCMC and MML. The MCMC had a better result than the MML method when the number of individuals were low. For the real data, MML time cost needed to reach convergence for some of the items was very expensive. However, the MML method is likely to be more effective to estimate the item parameters for greater individual sample sizes, but more potent computers and softwares are required to process the data.

By comparing the 2PL and Rasch LPE models, it was noticed that items with greater

values of a or λ have a steeper ICC and greater impact on the latent trait estimate. The 2PL item parameters are simpler and easier to estimate than “Rasch" LPE’s. Additionally, the former was a better fit for the General Science test. However, the LPE model had just a slightly worse fit in that case and the θ estimates under this model are affected by the the b parameter values, permitting a more logical ranking of the latent traits that can be useful in a classification test. Moreover, the “Rasch" LPE item parameter estimates could improve in a high-stakes test scenarios, where a large individual sample size is available.

Then, the LPE among with 2PL and Rasch models were applied to MST designs. Because of the model asymmetric nature, an MST based on it could be more likely to present items that are too easy or too difficult (since the maximum Fisher information occurs at probability of correct response 50%) on later stages of the test depending of the module assembly procedure. For that reason, three different module assembly methods based on Fisher, KL informations and CEM were implemented to analyze their performance under the LPE-MST.

The methodology were applied to simulated and a real datasets. The simulated data consisted in 100 “Rasch" LPE items and 5000 vector of responses. The MSTs for these cases consisted in a panel with a 1-3-3 structure. Guessing and accidental mistakes influences on the 2PL and “Rasch" LPE MST latent trait estimation were studied based on this data.

The General Science test from the Tetralogue project were used as the real data set, with 36 items and 1565 individuals and a MST structure was 1-2.

The Expected a Posteriori latent trait estimates were analyzed to evaluate which assembly method had better results based on Bias and RSME. The average probability of correct response was also included for comparison between methods.

For the simulated data, in relation to the assembly methods, the CEM had slightly better performance than the other 2 in both RSME and the average of probability being closer to 50%. Considering different models, although the 2PL model produced slightly biased estimates for the 100 items test, its multistage test had similar performance to the “Rasch" LPE model. One possible reason is that the “Rasch LPE" multistage test requires more items.

In relation to the guessing and mistake studies, the LPE performance, compared with 2PL, heavily depends on the items available. Since, the λ parameters can be stratified according to the stage or difficulty of modules, the LPE can be more lenient or punishing in these perspectives. If the high performance individuals guesses less than the low performance’s, it can be interesting to have items with high and low λ values on the hard and easy module, respectively. However, further analysis are necessary in that regard.

The GS data also pointed out the necessity of longer test for “Rasch" LPE item types. It is also important to mention that the “Rasch" LPE estimates had convergence problems because the item estimation requires a larger sample size.

In terms of future studies, further experiments for “Rasch" LPE (or LPE model) item

and latent trait estimation are necessary and MST could be explored in many different ways. Studies about item estimation with larger samples for the LPE family are necessary and other ways to simulate guessing and accidental mistakes can be done. Other MST scenarios with different structure (multiple panels, more stages and modules), assembly methods (for example: procedures that include restriction of probability of correct response being within a certain range could be implemented also), routing criteria and larger item sample availability could be implemented.

BIBLIOGRAPHY

ACHEN, C. H. Toward a new political methodology: Microfoundations and art. **Annual Review of Political Science**, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 5, n. 1, p. 423–450, 2002. Cited on page 2.

AZZALINI, A. The skew-normal distribution and related multivariate families. **Scandinavian Journal of Statistics**, Wiley Online Library, v. 32, n. 2, p. 159–188, 2005. Cited on page 2.

BARTON, M. A.; LORD, F. M. An upper asymptote for the three-parameter logistic item-response model*. **ETS Research Report Series**, v. 1981, n. 1, p. i–8, 1981. ISSN 2330-8516. Disponível em: <http://dx.doi.org/10.1002/j.2333-8504.1981.tb01255.x>. Cited on page 2.

BAZALDUA, A. D. L.; KHAN, S.; DAVIER, A. A. von; HAO, J.; LIU, L.; WANG, Z. On convergence of cognitive and noncognitive behavior in collaborative activity. In: . [S.l.]: Paper presented at The 8th International Conference on Educational Data Mining (EDM 2015), Madrid, 2015. Cited on page 34.

BAZÁN, J. L.; BOLFARINE, H.; BRANCO, M. D. *et al.* **A new family of asymmetric models for item response theory: A Skew-Normal IRT Family**. [S.l.]: Citeseer, 2004. Cited on page 2.

BOCK, R. D.; AITKIN, M. Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. **Psychometrika**, Springer-Verlag, v. 46, p. 443–459, 1981. ISSN 0033-3123. Disponível em: <http://dx.doi.org/10.1007/BF02293801>. Cited on page 11.

BOCK, R. D.; GIBBONS, R.; MURAKI, E. Full-information item factor analysis. **Applied Psychological Measurement**, v. 12, n. 3, p. 261–280, 1988. Disponível em: <http://apm.sagepub.com/content/12/3/261.abstract>. Cited on page 11.

BOCK, R. D.; LIEBERMAN, M. Fitting a response model forn dichotomously scored items. **Psychometrika**, Springer-Verlag, v. 35, n. 2, p. 179–197, 1970. ISSN 0033-3123. Disponível em: <http://dx.doi.org/10.1007/BF02291262>. Cited on page 13.

BOLFARINE, H.; BAZAN, J. L. Bayesian estimation of the logistic positive exponent irt model. **Journal of Educational and Behavioral Statistics**, v. 35, n. 6, p. 693–713, 2010. Disponível em: <http://jeb.sagepub.com/content/35/6/693.abstract>. Cited on page 2.

____. ____. **Journal of Educational and Behavioral Statistics**, v. 35, n. 6, p. 693–713, 2010. Disponível em: <http://jeb.sagepub.com/content/35/6/693.abstract>. Cited on page 31.

BOLT, D. M.; DENG, S.; LEE, S. Irt model misspecification and measurement of growth in vertical scaling. **Journal of Educational Measurement**, v. 51, n. 2, p. 141–162, 2014. ISSN 1745-3984. Disponível em: <http://dx.doi.org/10.1111/jedm.12039>. Cited on page 3.

BROOKS, S. P. Discussion on the paper by spiegelhalter, d. j., best, n. g., carlin, b. p., and van der linde, a. **Journal of the Royal Statistical Society, Series B**, v. 64, p. 616–618, 2002. Cited on page 3.

CHANG, H.; YING, Z. A global information approach to computerized adaptive testing. **Applied Psychological Measurement**, v. 20, n. 3, p. 213–229, 1996. Cited on page 20.

CHENG, Y. When cognitive diagnosis meets computerized adaptive testing: Cd-cat. **Psychometrika**, v. 74, n. 4, p. 619–632, 2009. ISSN 1860-0980. Disponível em: <http://dx.doi.org/10.1007/s11336-009-9123-2>. Cited on page 4.

ETS. **The GRE revised General Test**. 2015. <http://www.ets.org/ets/>. Accessed: 2015-6-2. Cited on page 1.

____. **Test of English as a Foreign Language (TOEFL)**. 2015. <http://www.ets.org/toefl/>. Accessed: 2015-6-2. Cited on page 1.

FLORES, S. E. A. **Modelos testest logíticos y logísticos de exponent positivo para pruebas de compresión de textos**. Dissertação (Mestrado) — Pontificia Universidad Católica del Perú, 2012. Cited 2 times on pages 2 and 31.

GELMAN XIAO-LI MENG, H. S. A. Posterior predictive assessment of model fitness via realized discrepancies. **Statistica Sinica**, Institute of Statistical Science, Academia Sinica, v. 6, n. 4, p. 733–760, 1996. ISSN 10170405, 19968507. Disponível em: <http://www.jstor.org/stable/24306036>. Cited on page 23.

HAMBLETON, R. K.; SWAMINATHAN, H.; ROGERS, H. J. **Fundamentals of Item Response Theory**. [S.l.]: Newbury Park, CA: Sage Publications, 1991. Cited on page 1.

INEP. **National High School Exam**. 2016. <http://www.portal.inep.gov.br/>. Accessed: 2016-1-11. Cited on page 2.

LINDEN, W. J. Van der; GLAS, G. A. W. **Computerized Adaptive Testing: Theory and Practice**. [S.l.]: Springer Science+Business Media, LLC, 233 Spring Street, New York, USA, 2000. Cited on page 3.

LINDEN, W. J. Van der; HAMBLETON, R. K. **Handbook of Modern Item Response Theory.** [S.l.]: Springer Science+Business Media, LLC, 233 Spring Street, New York, USA, 1997. Cited on page 1.

LINK, W. A.; EATON, M. J. On thinning of chains in mcmc. **Methods in Ecology and Evolution**, Blackwell Publishing Ltd, v. 3, n. 1, p. 112–115, 2012. ISSN 2041-210X. Disponível em: <http://dx.doi.org/10.1111/j.2041-210X.2011.00131.x>. Cited on page 38.

LORD, F. M. Some test theory for tailored testing. in w. h. holtzman (ed.). In: **Computer assisted instruction, testing, and guidance.** [S.l.]: New York, NY: Harper & Row, 1970. Cited on page 4.

MOLENAAR, D. Heteroscedastic latent trait models for dichotomous data. **Psychometrika**, Springer, v. 80, n. 3, p. 625–644, 2015. Cited on page 2.

MOLENAAR, D.; DOLAN, C. V.; BOECK, P. de. The heteroscedastic graded response model with a skewed latent trait: Testing statistical and substantive hypotheses related to skewed item category functions. **Psychometrika**, v. 77, n. 3, p. 455–478, 2012. ISSN 1860-0980. Disponível em: <http://dx.doi.org/10.1007/s11336-012-9273-5>. Cited on page 2.

RIGDON, S. E.; TSUTAKAWA, R. K. Parameter estimation in latent trait models. **Psychometrika**, Springer-Verlag, v. 48, p. 567–574, 1983. ISSN 0033-3123. Disponível em: <http://dx.doi.org/10.1007/BF02293880>. Cited on page 11.

ROSENBLATT, M. Remarks on some non-parametric estimates of a density function. **The Annals of Mathematical Statistics**, v. 27, p. 832–837, 1956. Disponível em: <http://projecteuclid.org/euclid.aoms/1177728190>. Cited on page 39.

RUBIN, D. B. Bayesianly justifiable and relevant frequency calculations for the applied statistician. **Ann. Statist.**, The Institute of Mathematical Statistics, v. 12, n. 4, p. 1151–1172, 12 1984. Disponível em: <http://dx.doi.org/10.1214/aos/1176346785>. Cited on page 23.

RUNDGREN, C.-J.; RUNDGREN, S.-N. C.; TSENG, Y.-H.; LIN, P.-L.; CHANG, C.-Y. Are you slim? developing an instrument for civic scientific literacy measurement (slim) based on media coverage. **Public Understanding of Science**, v. 21, n. 6, p. 759–773, 2012. Disponível em: <http://pus.sagepub.com/content/21/6/759.abstract>. Cited 2 times on pages 34 and 53.

SAMEJIMA, F. Departure from normal assumptions: A promise for future psychometrics with substantive mathematical modeling. **Psychometrika**, Springer-Verlag, v. 62, n. 4, p. 471–493, 1997. ISSN 0033-3123. Disponível em: <http://dx.doi.org/10.1007/BF02294639>. Cited on page 7.

____. Logistic positive exponent family of models: Virtue of asymetric item charactesitics curves. **Psychometrika**, v. 65, p. 319–335, 2000. Cited 2 times on pages 2 and 7.

SANDS, W. A.; WATERS, B. K. A global information approach to computerized adaptive testing. **Applied Psychological Measurement**, v. 20, 1996. Cited on page 19.

SANTOS, V. L. F.; GAMERMAN, D.; SOARES, T. M. **Teoria de Resposta ao Item: uma abordagem generalizada das Curvas Caracteristicas dos Itens**. Tese (Doutorado) — Rio de Janeiro Federal University, 2012. Cited on page 2.

SHANNON, C. E. A mathematical theory of communication. **The Bell System Technical Journal**, v. 27, p. 379–423, 1984. Cited on page 20.

SINHARAY, S. Bayesian item fit analysis for unidimensional item response theory models. **British Journal of Mathematical and Statistical Psychology**, Blackwell Publishing Ltd, v. 59, n. 2, p. 429–449, 2006. ISSN 2044-8317. Disponível em: <http://dx.doi.org/10.1348/000711005X66888>. Cited on page 23.

SPIEGELHALTER, D. J.; BEST, N. G.; CARLIN, B. P.; LINDE, A. V. D. Bayesian measures of model complexity and fit. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, Blackwell Publishers, v. 64, n. 4, p. 583–639, 2002. ISSN 1467-9868. Disponível em: <http://dx.doi.org/10.1111/1467-9868.00353>. Cited on page 3.

THISSEN, D. Marginal maximum likelihood estimation for the one-parameter logistic model. **Psychometrika**, Springer-Verlag, v. 47, p. 175–186, 1982. ISSN 0033-3123. Disponível em: <http://dx.doi.org/10.1007/BF02296273>. Cited on page 11.

TIERNEY, L. Markov chains for exploring posterior distributions. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 22, n. 4, p. 1701–1728, 1994. Cited on page 13.

Veldkamp, D. B. P. Item pool design and maintenance for multistage testing. In: Yan, D.; Davier, A. A. von; Lewis, C. (Ed.). **Computerized Multistage Testing: Theory and Applications**. Chapman and Hall/CRC Press, 2014. Disponível em: <http://doc.utwente.nl/88756/>. Cited on page 17.

VIANNA, H. M. **Testes em educação**. [S.l.: s.n.], 1982. Cited on page 1.

WANG, C.; CHANG, H. Item selection in multidimensional computerized adaptive testing - gaining information from different angles. **Psychometrika**, v. 76, n. 3, p. 363–384, 2011. Cited on page 20.

WILSON, D. T.; WOOD, R.; GIBBONS, R. D. Testfact: Test scoring, item statistics, and item factor analysis (computer software). **Chicago: Scientific Software International, Inc**, 1991. Cited on page 11.

YAN, D.; DAVIER, A. A. von; LEWIS, C. **Computerized Multistage Testing: Theory and Applications.** [S.l.]: CRC Press, 2014. Cited 2 times on pages 1 and 17.