UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Statistics on the evaluation of education
June 2016, Vol. 9, No. 1, pp. 119–139

Some methodological challenges in bayesian item response modeling: the assessment of standardized tests [1]

Algunos retos metodológicos del modelamiento bayesiano de teoría de respuesta al ítem: la calificación de pruebas estandarizadas

Andrés Gutiérrez[a]
agutierrez@icfes.gov.co

Diego Fernando Lemus[b]
dlemus@icfes.gov.co

William Fernando Acero[c]
wacero@contratista.icfes.gov.co

Abstract

The Ministry of National Education of Colombia (MEN), is increasingly concerned about the continuous improvement on teaching processes of teachers of foreign languages and has commissioned the assessment of this process to the Instituto Colombiano para la Evaluación de la Educación (Icfes), in order to assess the knowledge of those teachers and ensure that they conforms to the quality standards of the Common European Framework of Reference for Languages (CEFR). The reduction of the estimation error is a key objective in the rating process of the tests implemented by the Icfes, in order to improve accuracy when evaluating parameters such as the difficulty of the test and the performance of those who take it. In this paper a brief comparison between the classical and the bayesian methodology of the Rasch model is provided, comparing the behavior and conservation of classical assumptions in the second methodology. This paper shows besides how the Bayesian methodology reproduces a decrease in the estimation error of the parameters of skill and difficulty against classic methodology.

***Keywords*:** Rasch model , Bayesian Methodology, Estimation error .

[1] Gutiérrez, A., Lemus, D., H., Acero, W. (2016) Some methodological challenges in bayesian item response modeling: the assessment of standardized tests. *Comunicaciones en Estadística*, **9**(1), 119-139.

[a] Director of Assessment, Icfes

[b] Deputy Director of Instrument Design, Division of Assessment, Icfes

[c] Statistician, Division of Assessment, Icfes

Resumen

El Ministerio de Educación Nacional de Colombia (MEN) se preocupa cada vez más por la mejora continua en los procesos de enseñanza de los docentes que enseñan una lengua extranjera, y ha encargado la evaluación de dicho proceso al Instituto Colombiano para la Evaluación de la Educación (Icfes), con el fin de evaluar los conocimientos de dichos docentes para que los mismos cumplan con los estándares establecidos en el Marco Común Europeo de Referencia (MCER). La reducción del error de estimación resulta ser un objetivo fundamental en la calificación de las pruebas implementadas por el Icfes, con el fin de mejorar la precisión al momento de evaluar parámetros como la dificultad del examen y el desempeño de quienes lo presentan. En este artículo se encontrará una breve comparación entre la metodología clásica del modelo de Rasch y la metodología bayesiana del mismo, comparando el comportamiento y conservación de los supuestos clásicos en la segunda metodología. Además se muestra cómo la metodología bayesiana reproduce una disminución en el error de estimación de los parámetros de habilidad y dificultad frente a la metodología clásica.

***Palabras clave*:** modelo de Rasch, metodología bayesiana, error de estimación .

1 Introduction

In today's society, it is of great interest to constantly evaluate the learning of individuals. To this end, educational institutions, industries, and even the national government apply tests to determine the progress of the cognitive process of the educational system. Under this scenario, it's of vital importance to develop standardized tests that get closer and closer to what is pretended to evaluate. The key objective of Icfes is to offer the service of evaluation of education in all its levels and research about the factors affecting the education quality, in order to provide information to improve the education quality.

The MEN with the aim of defininf a solid and coherent system of evaluation has established clear lines to identify the training needs of teachers, the formulation of training plans, and in general, the close monitoring of the teaching and learning process of English in the country. Under that premise, the MEN adopted the Common European Framework of Reference for learning, teaching and assessment of English as foreign language since 2004 and implemented the National Bilingualism Program as strategy oriented to raise competition in English language in the national scope.

Since 2008, the MEN has been implementing diagnostic tests for English level of teachers, on an annual basis and in order to have updated and actual information on English levels of teachers of the Colombian official sector.The experience gained during the 40 years of the Icfes in the design and execution of various types of evaluation has allowed it to develop the technical and operational capacity for making other evaluations commissioned by public or private entities and derived incomes from them, as established by Law 635 of 2000. Therefore, since 2015, the Icfes under the direction of MEN assessed the level of use of English of the graduated teachers in this language, according to the Common European Framework of Reference for Languages: learning, teaching, and assessment-CEFR. English test requested evaluated three language skills through three components as described below:

> Listening component: This component has 30 questions to be answered in 40 minutes and allows to classify the evaluated in one of the following levels of the Common European Framework through five different parts A1 or less, A2, B1, B2 or higher.
>
> Reading component: This component has 45 questions to be answered in 60 minutes and allows to classify the evaluated in one of the following levels of the Common European Framework: Less than A1, A1, A2, B1, B2 or higher.
>
> Writing component: This component contains two parts to be answered in 45 minutes. The evaluated should make two writings, one short and another extensive in respond to a given context and given conditions. The two sides allow classifying the evaluated in one of the following levels of the Common European Framework (A1 or lower, A2, B1, B2 or higher).

This paper introduces the results of the scoring process of the listening component of the English test application in 2015 (implemented for a group of 8950 teachers graduated in English) using both the classical approach as the Bayesian approach of the item response theory. This research project is framed within one of the objectives of the Evaluation Office of Icfes consist in improving the score processes currently available, through the implementation of new methodologies that allow obtaining estimates of the difficulties of the items and skills of people, more precise and with less error.

To do this, Bayesian statistics gives an added value to current processes since, when having previous test data, it can be used to obtain better results in the score processes. This methodology could be adopted in the future as part of the statistical processing of tests.

This document has the following structure: In the next section we introduce the methodological frame considered to develop the research project; in section 3 we present the proposed methodology and in the forth section the obtained results. Finally, we present the conclusions of the study.

2 Methodologycal frame

Fox (2010) states that the models for the item response theory (IRT) were developed between 1970 and 1980, mainly from the field of psychometrics, trying to assess latent traits (such as the ability of the person), to obtain greater certainty of the conclusions that could be drawn from their studies. Some of the models that emerged in this period are Rasch model, one parameter model (1 PL model), two parameters (2 PL model), among others. In Sinharay (2003) it is stated that the computational constraint was a key factor since the methods of estimation of these models require a significant computational burden, which could not be supported by the developments of the time because the increasing complexity of the situations in which response data are collected poses new problems. Works like the Mislevy (1986), Rigdon & Tsutakawa (1983) and Swaminathan & Gifford (1982) have Bayesian extensions of the traditional item response models.

According to Ayala (2008), the logistic model of one parameter proposed by Rasch (1980) is one of the more used in item response theory (IRT) [1], , since its computational costs are low and it produces very good results. Under the logistic model of one parameter, the probability of answering an item correctly is mathematically defined as:

$$P(Y_{ik}=1|\theta_i,b_k)=\frac{e^{(\theta_i-b_k)}}{1+e^{(\theta_i-b_k)}}=(1+e^{(b_k-\theta_i)})^{-1} \tag{1}$$

Where θ_i is the individual's skill i, $i=1,\cdots,n$, y b_k is the item's difficulty k, $k=1,\cdots,J$. In other words under the Rasch model the probability for an individual to answer right depends on its ability and the difficulty of the item. As expected, as the ability of a person increases, the probability of answering right an item of b_k difficulty also, as shown in1.

The Bayesian approach of IRT uses computationally improved methods for modeling the discrete nature of data in the item response theory and generate so a new more flexible point of view that deals with the relationships with top-level data where standard distribution assumptions do not apply. A key element was the development of MCMC methods (Monte Carlo Markov Chains) and its simplicity for joint estimation despite the increase in model complexity. Specific problems related to response data modeling make some Bayesian methods very useful.

[1] It is understood as item to the question of a *test*.

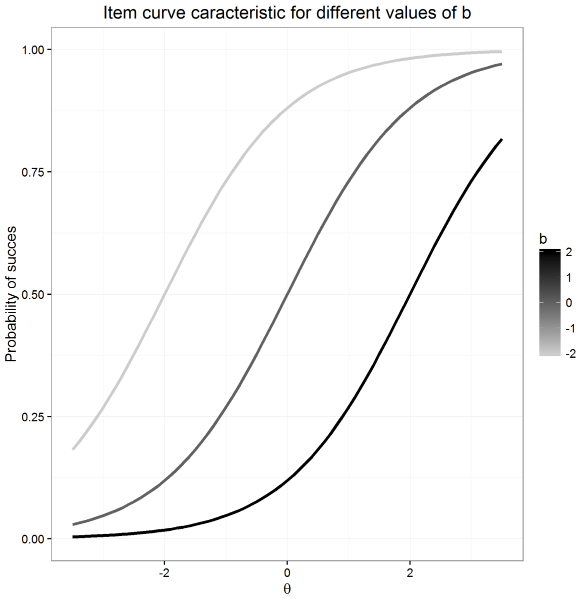

Figure 1: *Characteristic curve of the the item for different bk. Source: own elaboration.*

According to Fox (2010) the individual's ability θ and the difficulty of a b_k item are considered random variables in the Bayesian TRI and these parameters a prior distribution reflecting the uncertainty about the true values of these parameters before obtaining the data. The item response models discussed for the observed data describe the process of generating data as a function of unknown parameters in which are known as probability models. This is the part of the model that has the density of the conditional data in the model parameters. Therefore, there are two steps in the modeling process: in the first the specification of a distribution a priori is performed and in the second the specification of a probability model. In this case, we can assume prior distributions for the parameters as follows:

$$\theta \sim Normal(\mu_1, \tau_1^2) \tag{2}$$

$$b \sim Normal(\mu_2, \tau_2^2) \tag{3}$$

Where the vector of hiperparamenters is defined by $(\mu_1, \tau_1^2, \mu_2, \tau_2^2)$. In particular these prior distributions proposed are viable, since the θ parameter as the b parameter are in the interva $(-\infty, \infty)$, besides each $Y_{ik} \sim Bernoulli(p_{ik})$, where p_{ik} is the probability of answering right the $k - simal$ item. Suppose we have N independent realizations (N individuals presenting a test) for each of the k items of a test, which in turn measure a single construct independently. Therefore the verisimilitude function of the data is given by:

$$f(Y, b, \theta) = \prod_{i=1}^{N} \prod_{k=1}^{K} p_{ik}^{y_{ik}} (1 - p_{ik})^{y_{ik}} \tag{4}$$

We have besides that each p_{ik} is given by the equation 1 and replacing 1 in 4 we obtain:

$$f(Y, b, \theta) = \prod_{i=1}^{N} \prod_{k=1}^{K} e^{(\theta_i - b_k) y_{ik}} (1 - e^{(\theta_i - b_k)})^{y_{ik}} \tag{5}$$

Taken into account that the estimation of ability is affected by the estimation of difficulty, as shown in the verisimilitude, it is appropriate to use methods of Bayesian estimation to generate more reliable statistical inferences about the difficulties of the items (Fox 2010). To this end, it can be considered the information obtained in previous applications of the same test; after observing the data, information of the distribution a priori is combined with the information obtained in this application to generate densities a posteriori that allow making direct inference on parameters of interest. Flexibility in the definition of IRT models for the parameters of interest makes it possible to handle, for example, more complex sampling designs involving complex dependency structures and it's one of the strengths of the Bayesian approach.

3 Proposed procedure

As mentioned in the previous section, in the Bayesian approach, the model parameters are random variables and have a priori distribution reflecting uncertainty about the true values of these parameters before having observed data. In this regard there are two key things to consider: first, the specification of distributions a priori; and second, the verisimilitude function of the model, to make possible the combination of the two methods of estimation, i.e., Bayesian and classic. In this regard, the Bayesian inference on the parameters is performed under the conditional distributions of the posterior densities.

Moreover, recalling Bayes theorem, it is assumed that the data response is given by a latent variable θ and therefore, $p(\theta)$ represents the prior available information about these data and $p(y|\theta)$ makes reference to observed information of data; under this scheme it is possible to construct the following relationship:

$$p(\theta|y) \ \propto \ p(y|\theta)p(\theta) \tag{6}$$

Where $p(\theta|y)$ is the posterior distribution.

3.1 Example A

Initially it considered an example proposed by Fox (2010), in which it is supposed that a student with ability θ has the following vector of dichotomous responses $\mathbf{y}$ $= (1, 1, 0, 0, 0)^t$, where 1 indicates that the student responds the item right and 0 if not. The aim of the example is to estimate the posterior distribution for the parameter θ.

A particular case consist in assuming that all items have the same difficulty (for example, items of medium difficulty where $b_k = 0$). Under this scenario, the probit vesion of the Rasch where $P(Y_k = 1|\theta) = \Phi(\theta)$ defines the probability of answer correctly the *k-śimal* item. Then a prior uniform continuous distribution in the interval [-3,3] may be assumed for for θ such that $0.001 < \Phi(\theta) < 0.998$. The verisimilitude function $p(y|\theta)$ is given by:

$$p(y|\theta) = \Phi(\theta)^2(1 - \Phi(\theta))^3 \tag{7}$$

So, multiplying by the prior distribution $p(\theta) \propto 1$, the posterior distribution would be as follows:

$$p(\theta|y) \propto \Phi(\theta)^2(1 - \Phi(\theta))^3, \quad I(\theta)_{(-3,3)} \tag{8}$$

3.2 Example B

A more general scenario is to think that the difficulties of the items are not equal and are different from 0. Consider besides that n students have a test with k items. Under this scheme the prior distribution for the parameters of ability and difficulty are given by:

$$p(\theta, b|y) \propto p(y|\theta, b)p(\theta)p(b) \tag{9}$$

Where:

$$p(\theta) = \prod_n N(0, \sigma^2), \quad p(b) = \prod_k N(0, \tau^2) \tag{10}$$

With σ^2 and τ^2 hyperparameters of prior distributions. Whenever Gaussian distributions a priori are used the posterior distribution $p(\theta, b|y)$ doesn't have a known shape, whereby the approximations are necessary. In this case, some methods such as Laplacian are suitable.

3.3 Proposal

For the development of this work, a non-informative prior distribution for the parameter of ability and a non-informative prior distribution for the parameter of difficulty it were used. Also hiperparameters about the prior distribution were includes besides the difficulty parameter, in other words:

$$p(\theta) \sim N(0,1), \quad p(b) \sim N(\mu_1, \tau_1^2) \tag{11}$$

Where:

$$\mu_1 \sim N(0, 1/1000), \quad \tau_2 \sim Gamma(1/1000, 1/1000) \tag{12}$$

It is clear that the posterior distributions do not have a known shape, which, in this sense, computational part is an essential factor for the comparison of the two methods of estimation (classical and Bayesian) factor.

In the Bayesian case, the estimation process of parameters was made through the Gibbs [2], sampler using the program `JAGS`*(Just Another Gibbs Sampler)* [3] associated to program R. As initial values these parameters started at 0. It was chosen to perform 10000 simulations (draws) with a heating stage of 1000 draws.

4 Results

In this section obtained results in the scoring process of the listening component of the 2015 English test application (8950 teachers graduated in English) are introduced using both Bayesian and classical approach.

4.1 Estimation of the b and the θ

As only one simulation was made it was decided to use the diagnostic test of spectral density proposed by Geweke & Porter-Hudak (1983),from which for a total of 2030 parameters (2000 abilities and 30 difficulties) only one rate of non-convergence of 7.68% (figura 2) was obtained.

It is important to note that it's expected that as the difficulty of an item increases so does the ability of the person that answers that item too. Thus, it is expected that the more people respond to an item, the difficulty of this decrease proportionately. This relationship may be observed in figure 3. In addition, we can see that under the Bayesian approach the relationship between the percentages of correct answers of an item with the same difficulty remains, making clear that the item with the smaller difficulty is the one that has a lower percentage of correct answers. Although there is not much difference, it is worthwhile to establish which of these methodologies has less error of parameters estimation.

[2] http://halweb.uc3m.es/esp/Personal/personas/causin/esp/2012-2013/SMB/Tema8.pdf

[3] http://mcmc-jags.sourceforge.net/

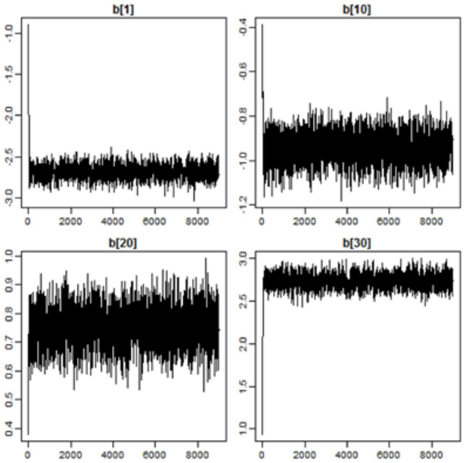

Figure 2: *Markov chains of some b. Source: own elaboration.*

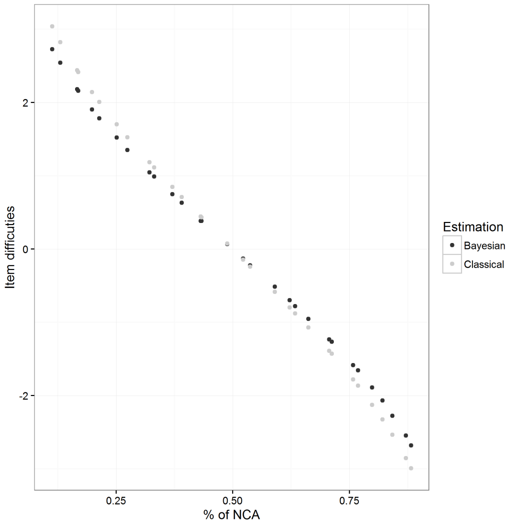

Figure 3: *Item difficulty vs percentage of right answers. Source: own elaboration.*

4.2 Comparison with the classical approach

For the Bayesian approach to be consistent, it must have at least a good relationship with the classical approach. In this sense, the correlation between the estimates should be high, though not necessarily close to one. In figure 4 you can see that as the estimated difficulty of an item grows in the classical approach it also increases in the Bayesian approach. Note that Bayesian approach maintains the assumption that difficulties of items are centered on zero, which is very good since it affirms that a different approach for these classical models works in a similar way.

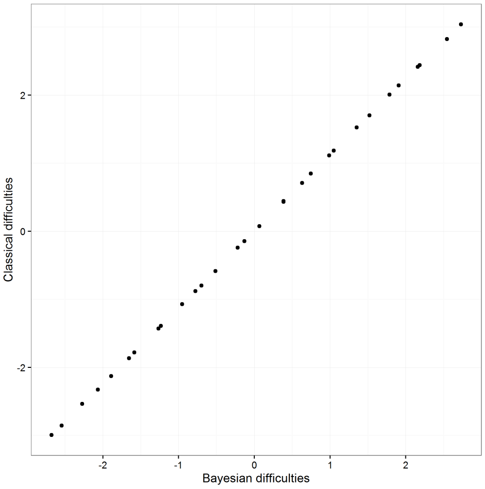

Figure 4: *Relationship between estimations. Source: own.*

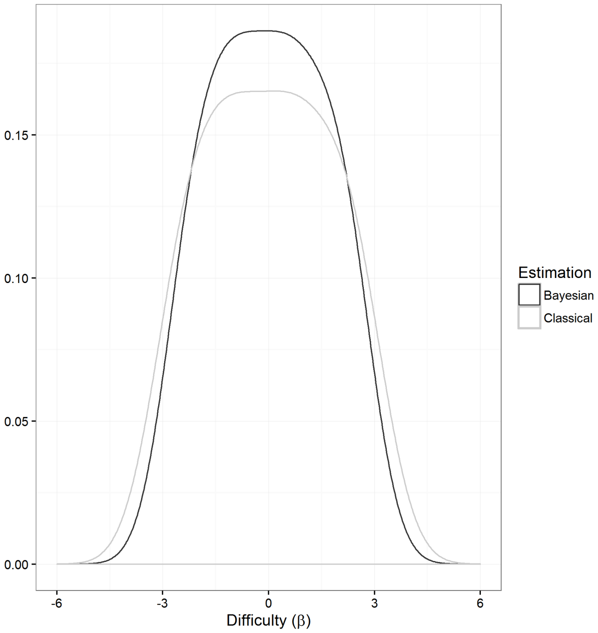

Figure 5: *Density of estimated difficulty. Source: own.*

Figures 5 and 6 show the density function of items and abilities of individuals. These functions are very similar, which was expected because of the above results. However as for the ability there is indo a notorious difference in bayesian estimation, above all because abilities are concentrating to a greater extent in the range [-2,2].

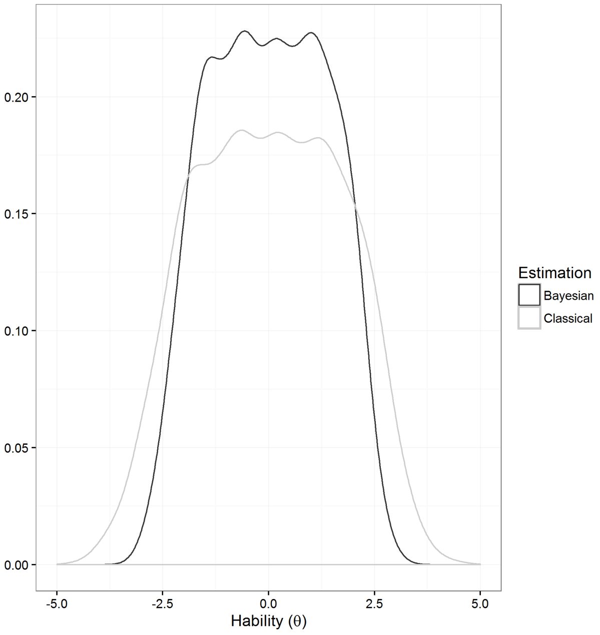

Figure 6: *Density of estimated ability. Source:own.*

Another important aspect to compare is the tendency of the difficulties of the items at the time of the simulation.When these data were simulated, it was set as a reference that the first item should have the lowest difficulty and the last one the highest. In figure 7 is evident that such behavior is maintained. It's quite particular that the estimate for items 14, 15 and 16 is the same in both methods, while the difficulty as it approaches the ends of the range, the difference in the estimated difficulty is much more marked.

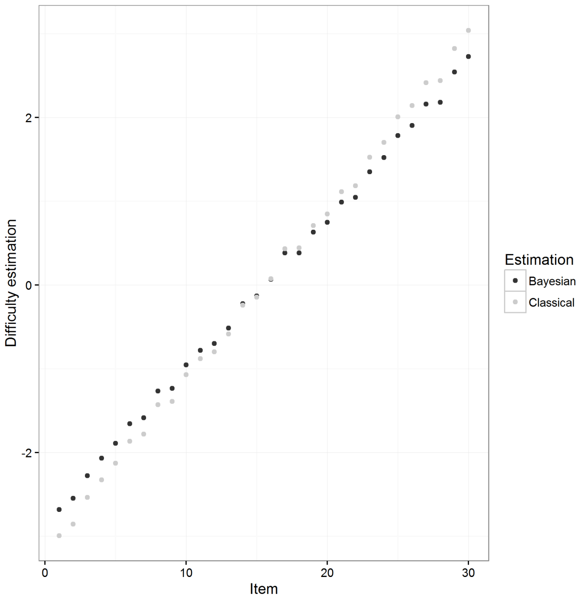

Figure 7: *Estimated difficulty of the item. Source: own.*

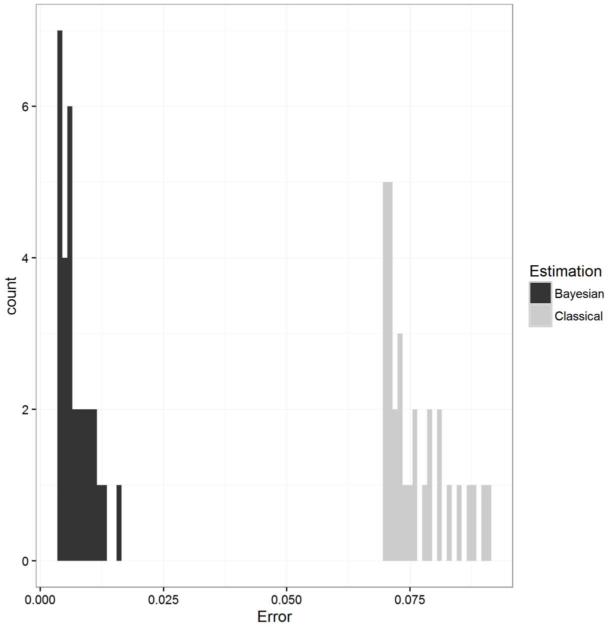

Figure 8: *Estimation error of difficulties. Source: own.*

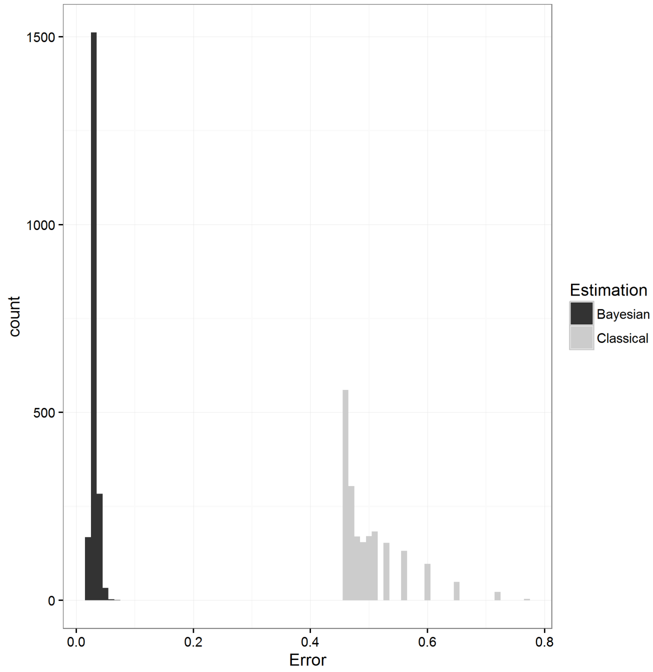

Figure 9: *Estimation error of abilities. Source: own*

Moreover, the process of identifying which of the two methodologies produces lower estimation error can be used as a criterion to define which of the two estimation processes is more accurate. Figures 8 and 9 show a quite notorious difference for abilities as for difficulties, where the Bayesian approach produces an estimation error much smaller. It's worthy to highlight that the distribution of the Bayesian estimation error keeps almost the same behavior in distributional terms as the classical methodology.

5 Conclusions

Bayesian methodology for estimating the parameters of difficulty and ability in a Rasch model proved to have a similar behavior to the classical methodology of such model, maintaining the assumption that both the abilities and the difficulties are centered on zero. Besides, it was found that the Bayesian methodology reduces the estimation error in the parameters of the Rasch model, implying a gain in precision regarding accuracy when scoring the tests.

Table 1: Estimation y estimation error of item difficulties

Bayesian			Classical		
Estimation	Error	Item	Mean	Error	Item
-2.6799	0.0134	1	-2.9860	0.0900	1
-2.5449	0.0160	2	-2.8460	0.0870	2
-2.2747	0.0108	3	-2.5280	0.0830	3
-2.0658	0.0104	4	-2.3210	0.0800	4
-1.8904	0.0093	5	-2.1190	0.0780	5
-1.6552	0.0079	6	-1.8600	0.0760	6
-1.5837	0.0071	7	-1.7740	0.0750	7
-1.2646	0.0063	8	-1.4250	0.0730	8
-1.2343	0.0062	9	-1.3840	0.0730	9
-0.9523	0.0060	10	-1.0630	0.0720	10
-0.7791	0.0060	11	-0.8710	0.0710	11
-0.6993	0.0048	12	-0.7900	0.0710	12
-0.5143	0.0048	13	-0.5780	0.0700	13
-0.2216	0.0040	14	-0.2360	0.0700	14
-0.1313	0.0046	15	-0.1390	0.0700	15
0.0672	0.0045	16	0.0800	0.0700	16
0.3848	0.0044	17	0.4380	0.0700	17
0.3847	0.0037	18	0.4480	0.0700	18
0.6313	0.0041	19	0.7140	0.0710	19
0.7471	0.0040	20	0.8510	0.0710	20
0.9885	0.0045	21	1.1180	0.0720	21
1.0479	0.0058	22	1.1880	0.0720	22
1.3517	0.0060	23	1.5300	0.0740	23
1.5203	0.0054	24	1.7060	0.0750	24
1.7854	0.0065	25	2.0120	0.0770	25
1.9061	0.0078	26	2.1460	0.0790	26
2.1585	0.0088	27	2.4200	0.0810	27
2.1827	0.0096	28	2.4450	0.0820	28
2.5424	0.0113	29	2.8280	0.0870	29
2.7278	0.0118	30	3.0410	0.0900	30

Another significant finding is that approximately 92% of chains converged, which reaffirms that estimates are feasible. Finally, although this methodology requires a greater amount of time and computing power, the advantage in terms of the reduction of error is enough to think about the implementation of this methodology to the classic scores processes currently performed by the Icfes.

Received: March 5, 2016
Accepted: April 18, 2016

References

Ayala, R. (2008), *The theory and pratice of item response theory*, 1 edn, The Guilford Press.

Fox, J. (2010), *Bayesian Item Response Modeling. Theory and Applications*, 1 edn, Springer.

Geweke, J. & Porter-Hudak, S. (1983), 'The estimation and application of long-memory times series models'.

Mislevy, R. (1986), 'Bayes model estimation in item response models', *Psychometrika* **51**(1), 177–195.

Rasch, G. (1980), *Probabilistic models for some intelligence and attainment tests*, 1 edn, University of Chicago Press.

Rigdon, S. & Tsutakawa, R. (1983), 'Parameter estimation in latent trait models', *Psychometrika* **48**(1), 567–574.

Sinharay, S. (2003), Bayesian item fit analysis for dichotomous item response theory models, Technical report, ETS, Princeton, NJ 08541.

Swaminathan, H. & Gifford, J. A. (1982), 'Bayesian estimation in the rasch model', *Journal of Educational Statistics* **7**(1), 175–192.

A Computational codes

```
rm(list = ls())
library(boot)
library(xtable)
library(ggplot2)
library(reshape2)
library(dplyr)
library(data.table)

p <- 30 # Number of items
n <- 2000 # Number of students

# student habilities
theta = seq(from = -3, to = 3, length.out = n)
# item difficulties
b <- seq(from = -3, to = 3, length.out = p)
# Matrix to create 1 and 0
pr <- y <- matrix(NA, nrow = n, ncol = p)

set.seed(11102015)
```

```
# Construction of responses according to Rasch model
for(i in 1:p){
    x <- theta - b[i]
    pr[, i] <- inv.logit(x)
    y[, i] <- rbinom(n, 1, pr[, i])
}

Rasch.data <- matrix(y, ncol = length(b))

PropNCA <- data.frame(Item = seq(from = 1, to = p, by = 1))
PropNCA[,"propOfNCA"] <- apply(Rasch.data, 2, mean)

# xtable(PropNCA, caption = "% Correct answers for Item",
 digits = 3)
GraphPropNCA <- ggplot() + geom_bar(data = PropNCA,
                     aes(y = propOfNCA, x = Item),
                     stat = "identity", fill = "black") +
                     ylab("% Correct answers") +
                     theme_bw()+ scale_colour_grey()
ggsave(plot = GraphPropNCA, filename = "../GraphPropNCA_ing.png")

# Student's NCA
PersoNCA <- data.frame(Persona = seq(from = 1, to = n, by = 1))
PersoNCA[,"NCA"] <- as.numeric(apply(Rasch.data, 1, sum))
#summary(PersoNCA[,"NCA"])
PropPerNCA <- prop.table(table(PersoNCA[,"NCA"]))*100
# xtable(PropPerNCA,
#  caption = "% of students according to number of correct answers")
PropPerNCA <- data.frame(PropPerNCA)
colnames(PropPerNCA) <- c("NCA","Estudiantes")

GraphEstuNCA <- ggplot() + geom_bar(data = PropPerNCA,
            aes(y = Estudiantes, x = NCA),
            stat = "identity", fill = "black") +
            ylab("% of Students") + theme_bw()+ scale_colour_grey()
ggsave(plot = GraphEstuNCA, filename = "../GraphEstuNCA_ing.png")


#########################################################
# # Classical estimation
#########################################################
library(mirt)
raschfit  <- mirt(Rasch.data, model = 1, itemtype='Rasch',SE = TRUE)

#########################
# Item difficulties
```

```
#########################
dificult  <- coef(raschfit, CI = 0.99, printSE = TRUE,
                           digits = 3, as.data.frame = TRUE)

filtroUno <- substr( rownames(dificult),8,8)
filtroDos <- substr( rownames(dificult),9,9)
dificult  <- subset(dificult, filtroUno == 'd' | filtroDos == 'd')
dificult  <- as.data.frame(dificult)
dificult[,"par"] <- dificult[,"par"] * -1 # multiplied by -1,
# because the result of mirt package is in the other side


Item <- seq(from =1, to = ncol(Rasch.data), by = 1)
Difficulty <- cbind(Item = Item, dificult)
names(Difficulty) <- c("Item","Difficulty","Error")
row.names(Difficulty) <- NULL
xtable(Difficulty)

# student habilities
scores <- fscores(raschfit, method = 'EAP', full.scores=TRUE,
                     full.scores.SE = TRUE) # Habilityes
scores <- data.frame(scores)
names(scores) <- c("Hability", "Error")
row.names(scores) <- NULL
xtable(scores[1:10,])

######################
# # %NCA vs Difficulty
######################
NCA <- apply(Rasch.data, 2, sum)
PorNCA <- NCA/n
NCADif <- cbind(Difficulty,PorNCA)
NCADifGra <- ggplot(data = NCADif) + geom_point(aes(x= Difficulty,
    y = PorNCA)) +  ylab("% NCA") + theme_bw()+ scale_colour_grey()
ggsave(plot = NCADifGra, filename = "../NCADifGra_ing.png")


#########################################################
# # Bayesian simulation of Rasch model
#########################################################

# Code WinBugs

library(R2jags)
library(coda)
library(lattice) # graphs
library(R2WinBUGS)
library(superdiag) # MCMC convergence criterion
library(mcmcplots) # graphs

# element definitions

Y <- Rasch.data # Matrix of 1's y 0's
burnin <- 100 # Burning
iter   <- 1000 # Simulations
chain  <- 1 # Chains
thin   <- 1 # Jumps to selection

Rasch.model <-function() {

    for (i in 1:n){
        for (j in 1:p){
            Y[i, j] ~ dbern(prob[i, j])
            logit(prob[i, j]) <-  ( theta[i] - b[j])
        }
        theta[ i ] ~ dnorm(0, 1)
    }

    for(j in 1:p){
        b[j] ~ dnorm(mu[j], tau[j])
        mu[j] ~ dnorm(0, 1/10000)
        tau[j] ~ dgamma(1/10000, 1/10000)
    }

}

Rasch.data  <- list("Y", "n", "p")
Rasch.param <- c("theta", "b")
Rasch.inits <- function(){list("theta"=rep(0,n), "b"=rep(0,p))}
set.seed(123)

Rasch.fit <- jags(data = Rasch.data, inits = Rasch.inits,
Rasch.param, n.chains = chain, n.iter = iter,
n.burnin = burnin, n.thin = thin, model.file = Rasch.model)

#########################################################
 # Convergence criteria
#########################################################
bayes.mod.fit.mcmc <- as.mcmc(Rasch.fit)
geweke <- geweke.diag(bayes.mod.fit.mcmc)
geweke <- data.frame(Z_Value = unlist(geweke),
               param = names(unlist(geweke)))
```

```
noconver <- which(geweke[,"Z_Value"] > qnorm(0.975) |
                 geweke[,"Z_Value"] < qnorm(0.025) )
NoConvergencia <- geweke[noconver,]
TasaNoConver <- nrow(NoConvergencia) / (nrow(geweke)-3)
TasaNoConver
# Gelman_Rubin <- gelman.diag(bayes.mod.fit.mcmc, transform=TRUE)
png("../Chain_betas_ing.png")
traplot(bayes.mod.fit.mcmc, parms=c("b[1]","b[10]","b[20]","b[30]"))
dev.off()
d <- summary(bayes.mod.fit.mcmc)
# superdiag(mcmcoutput = bayes.mod.fit.mcmc, burnin = 100 )

#########################################################
# # Parameters to be estimated
#########################################################
summary <- as.data.frame(d$statistics)
summary[,"Parameter"] <- row.names(summary)
summary[,"Parameter"] <- gsub("\\[|\\]","", summary[,"Parameter"])
betasEst    <- subset(summary, Parameter %like% "b")
thetasEst   <- subset(summary, Parameter %like% "theta")
devianceEst <- subset(summary, Parameter %like% "deviance")
#########################################################
# # Estimation concordances
#########################################################
propNCA <- apply(Y, 2, mean)
Verify <- data.frame(PropNCA = propNCA,
              Bayesian = betasEst[,"Mean"],
              Classical = dificult[,"par"])
colnames(Verify) <- c("PropNCA", "Bayesian", "Classical")

ggplot() + geom_point(data = Verify, aes(x=Bayesian, y = Classical),
                                            colour ="black") +
    xlab("Bayesian difficulties") +
    ylab("Classical difficulties") + theme_bw()+ scale_colour_grey()
ggsave(plot = last_plot(), filename = "../Verificacion_1_ing.png")

Verify <- melt(Verify, id.vars = "PropNCA")
names(Verify) <- c("PropNCA","Estimation", "Value")
ggplot() + geom_point(data = Verify,
                    aes(x=PropNCA, y = Value, colour = Estimation)) +
                 xlab("% of NCA") + ylab("Item difficuties") +
                 theme_bw()+ scale_colour_grey()
ggsave(plot = last_plot(), filename = "../Verificacion_2_ing.png")

#########################################################
# # Comparison methodologies
```

```
#########################################################
names(scores) <- c("Mean","Error")
scores[,"Persona"] <- seq(from = 1, to = nrow(scores), by = 1)
scores[,"Estimation"] <- "Classical"
thetasEst <- thetasEst[,c("Mean","Time-series SE","Parameter")]
row.names(thetasEst) <- NULL
colnames(thetasEst) <-c("Mean","Error","Persona")
thetasEst[,"Estimation"] <- "Bayesian"
thetas <- rbind(scores, thetasEst)
thetas[,"Persona"] <- gsub("theta","",thetas[,"Persona"])
row.names(thetas) <- NULL

# # Difficulty
dificult[,"Parameter"] <- seq(from = 1, to = nrow(dificult),by = 1)
colnames(dificult) <- c("Mean", "Error", "Item")
betasEst <- betasEst[,c("Mean","Time-series SE","Parameter")]
colnames(betasEst) <- c("Mean", "Error", "Item")
betasEst[,"Estimation"] <- "Bayesian"
dificult[,"Item"] <- gsub("b", "",dificult[,"Item"])
dificult[,"Estimation"] <- "Classical"
betas <- rbind(betasEst, dificult)


# Hability densities
ggplot() + geom_density(data = thetas, aes(x =  Mean,
         colour = Estimation ))+ xlab(expression(paste("Hability
         (",theta,")"))) + ylab("")+
                    xlim(-5,5) + theme_bw()+ scale_colour_grey()

ggsave(plot = last_plot(),filename = "../Densidad_Hability_ing.png")

# Difficulty densities
ggplot() + geom_density(data = betas, aes(x =  Mean,
                     colour = Estimation ))+ xlab(expression(paste
                     ("Difficulty (",beta,")"))) + ylab("")+
                     xlim(-6,6) + theme_bw()+ scale_colour_grey()

ggsave(plot = last_plot(),filename = "../Densidad_Difficulty_ing.png")

# Error estimation habilities
thetas %>%
    ggplot() +
    geom_histogram(aes(x=Error, fill = Estimation ),binwidth = 0.01)+
    theme_bw()+ scale_fill_grey()
ggsave(plot = last_plot(), filename = "../Error_Habilityes_ing.png")
```

```
# Error estimation difficulties
betas %>%
    ggplot() + geom_histogram(aes(x=Error, fill = Estimation ),
    binwidth = 0.001)+ theme_bw()+ scale_fill_grey()
ggsave(plot=last_plot(),filename = "../Error_Difficultyes_ing.png")

# Error estimation difficulties
betas[,"Item"] <- gsub("b","",betas[,"Item"])
betas %>%
    mutate(Item = as.numeric(Item)) %>%
    arrange(Item) %>%
    ggplot() +geom_point(aes(x=Item, y=Mean, colour = Estimation))+
    ylab("Difficulty estimation") + theme_bw()+ scale_colour_grey()
ggsave(plot=last_plot(), filename = "../Difficulties_Item_ing.png")

xtable(cbind(betas[1:30,-4],betas[31:60,-4]),digits = 4)
```