Revista Colombiana de Estadística
Edición para autor(es)

Bayesian Multi-faceted IRT models for measuring professor's performance in the classroom

Un modelo TRI de múltiples facetas bayesiano para la evaluación del desempeño docente en el aula

Karen Rosana Cordoba[1,a], Alvaro Mauricio Montenegro[1,b]

[1]Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia

Abstract

Evaluations of professor performance are based on the assumption that students learn more from highly qualified professors and the fact that students observe professor performance in the classroom. However, many studies question the methodologies used for such measurements, in general, because the averages of categorical responses make little statistical sense. In this paper, we propose Bayesian multi-faceted item response theory models to measure teaching performance. The basic model takes into account effects associated with the severity of the students responding to the survey, and the courses that are evaluated. The basic model proposed in this work is applied to a data set obtained from a survey of perception of professor performance conducted by Science Faculty of the Universidad Nacional de Colombia to its students. professor scores that are obtained as model outputs are real numerical values that can be used to calculate common statistics in professor evaluation. In this case, the statistics are mathematically consistent. Some of them are shown to illustrate the usefulness of the model.

***Key words*:** Multi-faceted IRT model, professor performance, Bayesian inference.

Resumen

Las evaluaciones del desempeño del profesor se basan en el supuesto de que los estudiantes aprenden más de profesores altamente calificados y el hecho de que los estudiantes observan el desempeño del profesor en el aula. Sin embargo, muchos estudios cuestionan las metodologías utilizadas para tales mediciones, en general, porque los promedios de las respuestas categóricas tienen poco sentido estadístico. En este artículo, proponemos modelos Bayesianos de Teoría de Respuesta al Ítem de múltiples facetas para medir

[a]Statistician. E-mail: krcordobap@unal.edu.co
[b]Associate professor. E-mail: ammontenegro@unal.edu.co

el desempeño. El modelo propuesto tiene en cuenta los efectos asociados con la severidad de los estudiantes que responden a la encuesta y los cursos que se evalúan. El modelo se aplica a un conjunto de datos obtenido de una encuesta de percepción del desempeño del profesor realizada por la Facultad de Ciencias de la Universidad Nacional de Colombia a sus estudiantes. Los puntajes del profesor que se obtienen como resultados del modelo son valores numéricos reales que se pueden usar para calcular estadísticas comunes en la evaluación del profesor. En este caso, las estadísticas son matemáticamente consistentes. Se muestra que algunos de ellos ilustran la utilidad del modelo.

***Palabras clave*:** Modelo TRI de múltiples facetas, Desempeño del profesor, Inferencia bayesiana.

1. Introduction

Studies evaluating professor performance have become an important aspect for administrative decision making in higher education institutions. Becker & Watts (1999) were the first to report that student evaluations in professor education were the most used method, and the only one in some cases.

One of the most important professor evaluation tools is the Student Evaluation of Teaching (SET). SET has some of the following objectives: to provide a diagnosis for the faculties about the performance of their professors; obtain measures of professors' effectiveness in order to make decisions about the institution staff and give information to students for the selection of courses and professors (Marsh 2007).

SET surveys usually include open and closed questions about professor performance in the course. An important characteristic of SET is the use of Likert-scales throughout the questionnaire.

In the course of university studies, students are exposed to different teaching styles that provide different professors. Therefore, a hypothesis about the SET could be that students give better grades to the professors from whom they learned more. The validity of these evaluations is based on the fact that students observe the performance of professors in the classroom and will therefore respond sincerely when they are asked about it.

There is strong evidence that the students' answers to the *teaching performance* questions do not measure such construct.

In this paper we discuss the validity of professor evaluation instruments through the implementation of statistical models which take into account some factors such the severity/leniency of students, the difficulty effect that differentiate the courses and so on.

The paper is organized as follows. Section 2 is a review about the factors that affect the evaluation of professor performance. An introduction to the multifaceted item response theory models is given in section 3. The Bayesian proposed model to measure performance professors is introduced in section 4. Some tools for evaluation of model adequacy are introduced in section 5.

This paper is based on a master's thesis in research, to expand the information presented the reader can refer to Cordoba (2020).

2. Evaluation of professors

Evaluation of professor comes from the first years of the past century. In this section we review some of the factors which influence the results of professor performance evaluation using SET.

2.1. Historical issues

The first formal grading scale for a professor evaluation was published in 1915 (Spencer & Flyr 1992). In the 1920s several of the main universities of the United States introduced the evaluation procedures for professors by the students (Marsh 1987).

During the 1970s and 1980s Feldman (1978, 1979, 1983, 1987, 1989) published a series of works, related to the factors that influence the results of professor performance evaluations. On the other hand, in the 1980s, two meta-analyzes were performed; Cohen (1981) and Feldman (1989) sought to collect information related to the correlations between SET results and student learning.

Uttl, White & Gonzalez (2017) published an updated meta-analysis of these correlations. The document identifies that the meta-analyzes of the 1980s have numerous problems related to the location of the studies used and that none of them is replicable.

Stark & Freishtat (2014) reviewed the SET scores from a statistical perspective. In the article they mention important points related to the validity of using the SET scores. Additionally Braga, Paccagnella & Pellizzari (2014) evaluate the content of student evaluations in contrast to measures of professor effectiveness. The authors use student performance to estimate measures of effectiveness.

2.2. Variables that can influence the scoring of SET

In general, there are two positions related to the use of SET. Those in favor set out different reasons for its use: Wachtel (1998) refers first to the economic issue, since he argue that these types of evaluations are cheap and easy to implement, secondly, he claims that SETs give value to student opinions. Murray (2005) emphasizes that students are the only ones who can assess their perceptions of professors in the classroom.

Those who oppose its use argue that SETs are assessments that measure student satisfaction and, therefore, student satisfaction is influenced by external factors that are not related to the effectiveness of teaching professor. For example, they argue that if a student scored differently than expected, their SET score is likely to be low, relative to their low satisfaction (Uttl, Eche, Fast, Mathison, Valladares Montemayor & Raab 2012).

Studies in education are extensive. It has been argued that variables can produce specific changes in the grades given to professors by students. Topics such as teaching methods, gender bias, academic rank and experience, difficulty of the course, personality characteristics, professor's reputation, sense of humor and degree of indulgence (Bélanger & Longden 2009) are included. In general, these variables are associated with three groups: course, student and professor variables.

It is common that the characteristics of the course have an influence when it comes to providing the grades by the students. Some of those characteristics that may influence the scores are: electivity of course (Feldman 1978), class schedule (Centra 1993, Feldman 1978, Koushki & Kunh 1982), course level (Feldman 1978, Marsh 1987), class size (Feldman 1978) and thematic area (Feldman 1978, Centra & Creech 1976).

In other hand, it is commonly accepted that in performance evaluation the most important variables associated directly with professor are: professor rank and work experience (Centra & Creech 1976, Feldman 1983), professor performance based on their reputation (Perry, Niemi & Jones 1974), productivity in research (Feldman 1987) and gender (Basow & Silberg 1987, Martin 1984).

Finally, the main hypothesis about SET is that students give better grades to the professors from whom they learned more. The validity of these evaluations is based on the assumption that students observe the performance of professors in the classroom and will therefore respond sincerely when they are asked about it. However, there are factors that really affect this hypothesis: personality of the student (Abrami, Perry & Leventhal 1982), interest in the area of knowledge of the course (Feldman 1977), student's gender (Feldman 1977), emotional state at the end of the course (Small, Hollenbeck & Haley 1982) and expectation and leniency hypothesis Marsh (1987).

If the reader is interesting in more information about these topics, is recommended that reviews the content of the different papers referenced in this section.

3. Multi-Faceted Rasch models

Multi-faceted Rasch models (MFRM) refer to a class of item response models suitable for a simultaneous analysis of multiple variables potentially having an impact on assessment outcomes (Eckes 2015). MFRM incorporates more variables, or facets, than the two that are included in a classical item response model. The first comprehensive theoretical statement was done by Linacre (1989). Based on this seminal work, substantive applications of MFRM have appeared in the fields of language testing, educational and psychological measurement (Barkaoui 2014), social behavior and the health sciences (Engelhard 2002, Engelhard 2013), and many others.

However, some disadvantages have also been reported from the applications of MFRM, the most important being related to scoring. The difficulty associated with objectively scoring the answers to the items contributes to a decrease in the reliability of the scores. In practice, raters are required to score examinees using a

specific rubric. Nevertheless, raters may influence examinees' scores in a number of ways.

Raters introduce variability into the scores gives to examinees that is associated with characteristics of the raters and not with the performance of examinees. In terms of regression models, rater variability is an unwanted variance component because it obscures the construct being measured. To solve this problem, a variance component is included in the MFRM, which authors identify as the severity/leniency component. The main objective in the MFRM is to include in the linear predictor those facets that have an impact on the scores awarded to examinees.

Consider as an example an educational test about writing in the English language. Assuming that the relevant facets have been identified, such as the examinees, tasks, and raters, an MFRM may be expressed as follows (Eckes 2015):

$$\log\left[\frac{p_{nljk}}{p_{nljk-1}}\right] = \theta_n - \delta_l - \alpha_j - \tau_k,$$

where

p_{nljk} = probability of examinee n receiving a rating k from rater j on task l,
p_{nljk-1} = probability of examinee n receiving a rating $k-1$ from rater j on task l,
$\theta_n =$ ability of examinee n,
$\delta_l =$ difficulty of task l,
$\alpha_j =$ severity of rater j,
$\tau_k =$ difficulty of receiving a rating of k relative to $k-1$.

4. Bayesian Multi-faceted Model for Measuring Professor Performance

In this section the proposed Bayesian multi-faceted (BMF) model is introduced as a tool to be applied to SET data. However, applications of the model in other areas are not only possible but are welcomed. In particular, the BMF could be used in any application of the MFRM.

In addition, the ordered logistic distribution is introduced. The definitions of the BMF models introduced are based on this distribution.

4.1. Ordered Logistic Distribution

The inverse logit function is defined as $\text{logit}^{-1}(x) = (1+e^{-x})^{-1}$. This expression defines the cumulative distribution function (cdf) of the logistic distribution. Let $\boldsymbol{\beta}' = (\beta_1, \ldots, \beta_{K-1}) \in \mathbb{R}^{K-1}$, such that $\beta_k < \beta_{k+1}$, and let $\eta \in \mathbb{R}$. Let $K \in \mathbb{N}$

with $K > 2$. Then for $k \in \{1, \ldots, K\}$, the probabilistic mass function (pmf) of the ordered logistic distribution is defined as follows.

$$g(k|\eta, \boldsymbol{\beta}) = \begin{cases} 1 - \text{logit}^{-1}(\eta - \beta_1) & \text{if } k = 1, \\ \text{logit}^{-1}(\eta - \beta_{k-1}) - \text{logit}^{-1}(\eta - \beta_k) & \text{if } 1 < k < K, \\ \text{logit}^{-1}(\eta - \beta_{K-1}) & \text{if } k = K. \end{cases}$$

The $k = 1$ and $k = K$ edge cases can be subsumed into the general definition by setting $\beta_0 = -\infty$ and $\beta_K = \infty$ with $\text{logit}^{-1}(-\infty) = 0$ and $\text{logit}^{-1}(\infty) = 1$.

In a classical logistic regression, the η-values are known predictors, while the β_k-values are regression parameters to be estimated. In this case, the η's are latent variables to be predicted, and the β_k-values are item parameters. Note that η and the β_k's are in the same space. On the other side, for a fixed value of η, if $\beta_{k-1} < \eta \leq \beta_k$, then, the category k generally has the highest probability. Figure 1 illustrates this fact. Consequently, the β-parameters are cut points that determine the probability of each category depending on the value of η.

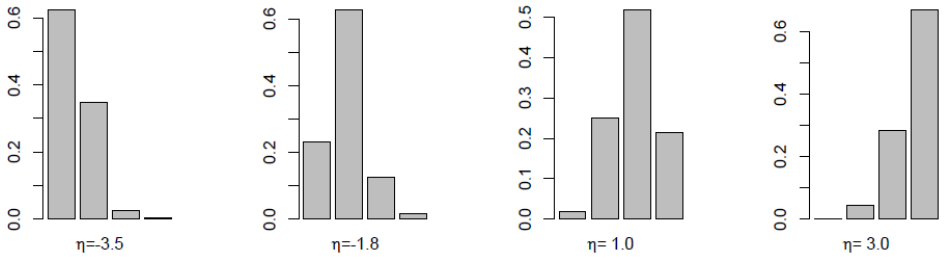

FIGURE 1: Probabilistic mass function of the Ordered Logistic distribution for different values of η, with $\boldsymbol{\beta} = (-3.0, 0.0, 2.3)^t, K = 4$.

4.2. Bayesian Multi-faceted Model for Measuring Professor Performance

It is assumed that professor performance is measured by using a latent variable. From a general statistical perspective, latent variables can be considered random effects (Bartholomew, Knott & Moustaki 2011). Moreover, the severity/leniency of the student when evaluating a professor is measured by a new latent variable. In each course, each student grades a professor in all items of the professor's performance questionnaire. It is assumed that each item is based on a Likert scale with K_j categories; the questionnaire has p items, and N professors are evaluated. The student selects a category of the item to grade the professor.

The Bayesian Multi-Faceted model is then defined as follows. Let θ_i be the latent variable which measures the performance of the $i-$th professor. Let $\boldsymbol{\gamma}_i' = (\gamma_{i1}, \cdots, \gamma_{i,n_i})$ be the vector of latent variables which measures the severity of the students evaluating the professor i. The value n_i is the total number of students evaluating the professor i. Let κ_c a random effect associated to the course c. Let

$\boldsymbol{\beta}_j' = (\beta_{j1}, \ldots, \beta_{jK_j-1})$ be cut points to the item j associated with each response category. Let Y_{ijsc} be the rating that the s-th student assigns to professor i for the item j in the course c. Thus, the conditional probability that $[Y_{ijsc} = k|\theta_i, \gamma_{is}, \boldsymbol{\beta}_j, \kappa_c]$ is given by

$$Prob[Y_{ijsc} = k|\theta_i, \gamma_{is}, \boldsymbol{\beta}_j, \kappa_c] = logit^{-1}(\eta_{isc} - \beta_{j(k-1)}) - logit^{-1}(\eta_{isc} - \beta_{jk}), \quad (1)$$

where $k = 1, \cdots, K_j$, $c = 1, \cdots, C$, $j = 1, \cdots, p$, $s = 1, \cdots, n_i$, $i = 1, \ldots, N$, and, $\eta_{isc} = \theta_i - \kappa_c - \gamma_{is}$. The model defined in the equation (1) is not identifiable as it is common in item response models. To set a scale, we assume that $\theta_i \sim N(0, 1)$. In addition, the following prior distributions are assigned.

$$\begin{aligned}
\kappa_c &\sim N(0, \sigma_\kappa^2), \\
\gamma_{is} &\sim N(0, \sigma_\gamma^2), \\
\beta_{jk} &\sim N(\mu_\beta, \sigma_\beta^2), \text{ cut points} \\
\mu_\beta &\sim N(0, 2), \\
\sigma_\kappa &\sim Cauchy(0, 2)I_{(0,\infty)}, \\
\sigma_\gamma &\sim Cauchy(0, 2)I_{(0,\infty)}, \\
\sigma_\beta &\sim Cauchy(0, 2)I_{(0,\infty)}.
\end{aligned}$$

To have the posterior distribution of the model, two assumptions are required. First, the responses of the students are independent. Second, the responses of the student to a professor are independent. These assumptions, especially the second, may be controversial. The first assumption may be violated if a student responds to more than one questionnaire for the same professor. This can occur if the student is taking two courses with the same professor. However, in this case this problem is diminished because the subjects are different. Furthermore, there is no way to know when this situation occurs because the students' responses are anonymous. On the other hand, each question in the questionnaire is designed to measure a different aspect of the professor's performance in the classroom. Therefore, it is expected that each one of the questions is answered independently by the student.

Let $p_{ijksc} = Prob[Y_{ijsc} = k|\theta_i, \gamma_{is}, \boldsymbol{\beta}_j, \kappa_c]$. Thus, the posterior distribution for the Bayesian multi-faceted model is given by

$$\begin{aligned}
L[\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{y}] \propto & \prod_{i=1}^{N} \prod_{j=1}^{p} \prod_{s=1}^{n_i} \prod_{c=1}^{C} \prod_{k=1}^{K_j} [p_{ijksc}]\chi_k(y_{ijsc}) \times \\
& p(\theta_i)p(\beta_{jk}|\mu_\beta, \sigma_\beta)p(\gamma_{is}|\sigma_\gamma)p(\mu_\beta)p(\kappa_c|\sigma_\kappa)p(\sigma_\beta)p(\sigma_\gamma)p(\sigma_\kappa), \quad (2)
\end{aligned}$$

Bold Greek letters represent the corresponding complete vector of parameters. Vector $\boldsymbol{y}$ is the complete vector of the observed responses, and $\chi_k(y_{ijsc})$ is defined as

$$\chi_k(y_{ijsc}) = \begin{cases} 1, \text{ if } y_{ijsc} = k. \\ 0, \text{ otherwise.} \end{cases} \quad (3)$$

5. Tools for evaluating model adequacy of BMF Models

To be sure that the BMF models are good to measure the professor's performance, it is necessary to guarantee:

1. **Unidimensionality:** The items in the questionnaire are designed to measure a unique construct.
2. **Independence:** The latent traits θ_i must be independent of the items and from the socio-demographic characteristics of the students, the students' severity, and their scores in the course.
3. **Goodness of fit (GoF):** It is necessary to assess the quality of the model to fit the data.

The BMF models proposed in this paper are unidimensional. That is, the items in the questionnaire are designed to measure a unique construct. In section 5.1, we introduce some tools to assess the unidimensionality of the data.

The second important issue is that the latent traits θ_i are independent from the items in the questionnaire and from the socio-demographic characteristics of the students, the students' severity, and their scores in the course. Furthermore, it is assumed that the experts who design the questionnaires guarantee independence between the questionnaire and professors. The theoretical aspects of previously discussed concerns are based on item response theory (Bock 1997, Baker & Kim 2004, Birnbaum 1968). Nevertheless, independence between the latent traits and the socio-demographic characteristics of the students, their severity and their score in the course must be assessed.

In the item response models, it is necessary to assess the goodness of fit of the model to the complete data, the adequacy of the model for fitting the data of each one of the examinees, and the data of each one of the items. In section 5.2, a GoF statistics is proposed to do that, and statistics to compare models are also presented.

5.1. Analysis of unidimensionality

The first principle to define a scale to measure a unique construct is that the resulting data be unidimensional.There are several tools to evaluate the unidimensionality of data. Three of them were used in this study:

First, the principal component analysis (PCA) as a dimensionality reduction technique. There are different interpretations about the dimensions of the data

based on the plot of the eigenvalues from PCA. In general, a first eigenvalue that is very large compared to the other values suggests unidimensionality (Jollife 2002).

Second, Cronbach's alpha reliability coefficient is an index between 0 and 1 based on the variance of the total score and the variance of the score of each item. The extreme value 0 means that the items are not correlated, and value 1, means that all the items are the same. According to the experts, values greater than 0.75 or 0.80 mean that the data is measuring a unique construct (Cronbach 1951, Lord & Novick 2013).

Finally, the Cronbach Mesbah Curve (CMC) shows Cronbach's alpha coefficient after the item for which the preserved data has a maximal Cronbach's alpha coefficient is removed. If the data is unidimensional, the curve is monotonically increasing. For further details see the work of Cameletti & Caviezel (2012).

5.2. Goodness of fit statistics

For ease, in this section, it is assumed that the complete vector parameter is $\boldsymbol{\theta}$, and the observed data is $\boldsymbol{y}$. The likelihood of the observations is $f(\boldsymbol{y}|\boldsymbol{\theta})$ and the prior density of $\boldsymbol{\theta}$ is $\pi(\boldsymbol{\theta})$; hence the model specification is $f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$. The predictive distribution of unobserved values $\boldsymbol{\omega}$ is denoted $f(\boldsymbol{\omega}|\boldsymbol{y})$, and defined as

$$f(\boldsymbol{\omega}|\boldsymbol{y}) = \int f(\boldsymbol{\omega}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}.$$

In the GoF procedures, the data $\boldsymbol{\omega}$ are named replicate data. Let $\boldsymbol{W}$ be the random variable with density $f(\boldsymbol{\omega}|\boldsymbol{y})$. For evaluating the adequacy of a model to the observed data, Box (1980) proposed to use the predictive distribution $f(\boldsymbol{\omega}|\boldsymbol{y})$. In particular, Box proposed to compute the expectation $E[g(\boldsymbol{W},\boldsymbol{\theta})|\boldsymbol{y}]$ of some relevant checking statistic $g(\boldsymbol{W},\boldsymbol{\theta})$. Gelfand, Dey & Chang (1992) proposed several checking functions based on discrepancy measures. In this work, a checking function based on the discrepancy statistic $D(\boldsymbol{W},\boldsymbol{\theta}) = -2\log f(\boldsymbol{W}|\boldsymbol{\theta})$ is proposed. The constant 2 is redundant, nevertheless, it was included in order to have a counterpart to the log likelihood statistics commonly used in frequentist statistics. The idea of using discrepancy measures is studied in deep by Gelman, Meng & Stern (1996).

According to the notation used by Box (1980) and Gelfand et al. (1992) and using the discrepancy statistic D as proposed by Gelman et al. (1996), we construct the goodness of fit statistics as follows. Define $B_{\omega,\theta} = \{(\boldsymbol{\omega},\boldsymbol{\theta}) : D(\boldsymbol{\omega},\boldsymbol{\theta}) \leq D(\boldsymbol{y},\boldsymbol{\theta})\}$. The check statistics is defined as $g(\boldsymbol{W},\boldsymbol{\theta}) = I_{B_{\omega,\theta}}(\boldsymbol{W},\boldsymbol{\theta})$. The statistical decision is based on the value $d_{\omega,\theta}$ defined as

$$\begin{aligned} d_{\omega,\theta} &= P(B_{\omega,\theta}) \\ &= E[g(\boldsymbol{W},\boldsymbol{\theta})|\boldsymbol{y}] \\ &= \int\int g(\boldsymbol{\omega},\boldsymbol{\theta})f(\boldsymbol{\omega}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}d\boldsymbol{\omega}. \end{aligned}$$

Values of $d_{\omega,\theta}$ around 0.5 confirm that the model fit well the data. A reasonable range can be between 0.05 and 0.95, or more strictly, between 0.1 and 0.9.

5.2.1. Goodness of fit for Bayesian Multi-faceted models

In item response models, it is common to assess the model fit to the complete data and to items and people separately. Let $\boldsymbol{y}_{i..}$ be the complete vector of responses to professor i, $\boldsymbol{y}_{.j.}$ be the complete vector of responses to item j. The corresponding replicate data will be denoted $\boldsymbol{\omega}_{i..}$, and $\boldsymbol{\omega}_{.j.}$ respectively. To assess the GoF of the model to the i-th professor's data, the predictive density is used, given by

$$f(\boldsymbol{\omega}_{i..}|\boldsymbol{y}_{i..}) = \int f(\boldsymbol{\omega}_{i..}|\theta_i, \boldsymbol{\gamma}_i, \boldsymbol{\kappa}_i, \boldsymbol{\beta})\pi(\theta_i, \boldsymbol{\gamma}_i, \boldsymbol{\kappa}_i, \boldsymbol{\beta}|\boldsymbol{y}_{i..})d\theta_i d\boldsymbol{\gamma}_i d\boldsymbol{\kappa}_i, d\boldsymbol{\beta}. \tag{4}$$

Similarly, the predictive density to assess the fit to the j-th item is given by

$$f(\boldsymbol{\omega}_{.j.}|\boldsymbol{y}_{.j.}) = \int f(\boldsymbol{\omega}_{.j.}|\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\kappa}, \boldsymbol{\beta}_j)\pi(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\kappa}, \boldsymbol{\beta}_j|\boldsymbol{y}_{.j.})d\boldsymbol{\theta} d\boldsymbol{\gamma} d\boldsymbol{\kappa} d\boldsymbol{\beta}_j. \tag{5}$$

Finally, the predictive density to assess the fit to the complete data test is given by

$$f(\boldsymbol{\omega}|\boldsymbol{y}) = \int f(\boldsymbol{\omega}|\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\kappa}, \boldsymbol{\beta})\pi(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\kappa}, \boldsymbol{\beta}|\boldsymbol{y})d\boldsymbol{\theta} d\boldsymbol{\gamma} d\boldsymbol{\kappa} d\boldsymbol{\beta}. \tag{6}$$

5.2.2. Computational approach to compute the GoF statistics

Let $(\boldsymbol{\theta}^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^{(t)}), \boldsymbol{\kappa}^{(t)})$ be the complete t-th sample from the posterior distribution $\pi(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\kappa}|\boldsymbol{y})$, with $t = 1, \cdots, T$. The data come from the estimation output produced by Stan. The GoF statistic for the data of i-th professor can be calculated as follows:

1. Obtain a replicate data vector $\{\boldsymbol{\omega}_{ijsc}; j = 1, \cdots, p; s = 1, \cdots, n_i\}$ from the ordered logistic distributions given by

$$\begin{aligned} g_{ijsc}(k|\theta_i^{(t)}, \kappa_c^{(t)}, \gamma_{is}^{(t)}, \boldsymbol{\beta}_j^{(t)}) = \text{logit}^{-1}(\theta_i^{(t)} - \kappa_c^{(t)} - \gamma_{is}^{(t)} - \beta_{j,k-1}^{(t)}) \\ - \text{logit}^{-1}(\theta_i^{(t)} - \gamma_{is}^{(t)} - \kappa_c^{(t)} - \beta_{j,k}^{(t)}), \end{aligned}$$

   $k = 1, \cdots, K_j$. Remember that $\beta_{j0} = -\infty$, and $\beta_{jK_j} = \infty$.

2. The discrepancy measure for $\boldsymbol{\omega}_{i..}$ is given by

$$D_i(\omega_{i..}|\theta_i^{(t)}, \kappa_c^{(t)}, \gamma_i^{(t)}, \beta_j^{(t)}) = -2\sum_{j=1}^{p}\sum_{s=1}^{n_i}\sum_{c=1}^{C}\sum_{k=1}^{K_j}\chi_k(\omega_{ijsc})\log g_{ijsc}(k|\theta_i^{(t)}, \gamma_{is}^{(t)}, \kappa_c^{(t)}, \beta_j^{(t)})$$

3. The discrepancy measure for $\boldsymbol{y}_{i..}$ is given by

$$D_i(y_{i..}|\theta_i^{(t)},\kappa_c^{(t)},\gamma_i^{(t)},\beta_j^{(t)}) = -2\sum_{j=1}^{p}\sum_{s=1}^{n_i}\sum_{c=1}^{C}\sum_{k=1}^{K_j}\chi_k(y_{ijsc})\log g_{ijsc}(k|\theta_i^{(t)},\gamma_{is}^{(t)},\kappa_c^{(t)},\beta_j^{(t)})$$

4. The GoF statistic is given by

$$p_i = \frac{1}{T}\sum_{t=1}^{T} 1_{(D_i(\boldsymbol{\omega}_{i..}|\theta_i^{(t)},\boldsymbol{\gamma}_i^{(t)},\boldsymbol{\kappa}_c^{(t)},\boldsymbol{\beta}^{(t)})-D_i(\boldsymbol{y}_{i..}|\theta_i^{(t)},\boldsymbol{\gamma}_i^{(t)},\boldsymbol{\kappa}_c^{(t)},\boldsymbol{\beta}^{(t)})>0)}.$$

Values of p_i close to 0.5 mean a good fit. The GoF statistic for the data of j-th item can be compute similarly.

5.3. Model selection criteria

There are a variety of procedures for the selection of models within the framework of Bayesian inference. In this work we used WAIC and LOO.

5.3.1. Watanabe information Criterium

The Watanabe information Criterium (WAIC) was introduced by Watanabe (2010). For discussion and details see Gelman, Hwang & Vehtari (2014) and Ariyo, Quintero, Muñoz, Verbeke & Lesaffre (2019). For a future observation $\tilde{y}_i$, this criterion measures the predictive accuracy of the model based on the log-posterior predictive distribution $\log p_{\theta|y}(\tilde{y}_i)$ of the vector of parameters $\boldsymbol{\theta}$. Since $\tilde{y}_i$ is unknown, predictive accuracy is defined by the expected log-predictive distribution (elpd) as

$$elpd_i = E_f[\log p_{\theta|y}(\tilde{y}_i)] = \log p_{\theta|y}(\tilde{y}_i)f(\tilde{y}_i)d\tilde{y}_i,$$

where f is the distribution unknown to the true model. For each observation of a new data set, *elpd* is calculated to establish the predictive accuracy of that data set. The punctual logarithmic predictive distribution (*lppd*) based on the observed data and is calculated as follows:

$$lppd = \log\prod_{i=1}^{n} p_{\theta|y_i}(y_i) = \sum_{i=1}^{n}\log\int_{\theta} p(y_i|\theta)p(\theta|y)d\theta$$

In practice we use a sample from the posterior distribution of the parameter to estimate *lppd* as

$$\widehat{lppd} = \sum_{i=1}^{n}\log\left[\frac{1}{K}\sum_{k=1}^{K} p(y_i|\theta^k)\right] \tag{7}$$

An estimation of the effective number of parameters p_{WAIC} is given by

$$p_{WAIC} = 2\sum_{i=1}^{n}\left[\log\left(\frac{1}{K}\sum_{k=1}^{K}p(y_i|\theta^k)\right) - \frac{1}{K}\sum_{k=1}^{K}\log p(y_i|\theta^k)\right] \tag{8}$$

Then WAIC is given by

$$WAIC = -2\widehat{lppd} + 2p_{WAIC} \tag{9}$$

5.3.2. Cross-validation Leave-one-out

Cross-validation is an approach to estimate predictive accuracy outside the sample using adjustments within the sample. It requires re-adjusting the model with different training sets. Cross-validation Leave-one-out (LOO) can be easily calculated using importance sampling (Gelfand et al. 1992). Vehtari, Gelman & Gabry (2017) showed the development of statistics. They consider the calculations using the log-likelihood evaluated in the usual subsequent simulations of the parameters defined in the equation (7). The Bayesian LOO estimate of the out-of-sample predictive adjustment is

$$elpd_{loo} = \sum_{i=1}^{n}\log p(y_i|y_{-i}), \tag{10}$$

where

$$p(y_i|y_{-i}) = \int p(y_i|\theta)p(\theta|y_{-i})d\theta \tag{11}$$

is the predictive density given the information without the i-th data point.

6. Evaluation of teaching performance at Universidad Nacional de Colombia

The data comes from a student survey that is conducted every semester at the Faculty of Science in Universidad Nacional de Colombia. The survey is designed to measure the perception of professor performance in the classroom, the relevance of the course and the efficiency of the resources available on campus. The questionnaire can be reviewed in Appendix A. It contained 25 questions in 8 sections as following:

1. **General student data:** Information about sociodemographic characteristics of students without requesting explicit personal identification.

2. **Course management:** Inquire about topics related to program fulfillment, regular professor attendance, and if group work is encouraged and the achievements of the students by the professor are recognized.

3. **Impact of the course:** Information on the impact of the subject on vocational training; topics related directly to the contents are directed.

4. **Learning environments:** Inquire about issues related to the use of other spaces for the development of the subject and the promotion of self-teaching by the professor.

5. **Use of technology:** Inquire about the use of digital technologies and tools for the development of skills by students.

6. **Physical plant and implements:** Information on the infrastructure conditions and the quality of the implements used for the development of the course.

7. **Auxiliary:** Inquire about the need and importance of the support of the auxiliary in the course.

8. **General:** Inquire about issues related to strengths and weaknesses of the general performance of the professor and the content of the subject.

Two Likert scales were used to code the responses. The first is coded as *Poor*, *Bad*, *Regular*, *Good* and *Excellent*. This scale is aimed at measuring the perception of the overall performance of the student himself, and some aspects of the students perception about the professor performance and the content of the subject. The second scale is code as *Never*, *Almost Never*, *Sometimes*, *Almost Always* and *Always* and is used in specific questions about the course management.

6.1. Preliminary analysis

Data from the second semester of 2015 were used. The initial data contains responses associated with 14703 records. The information is related to professor, course and student. In addition to this, data related to the department that offers the subject and the sex of professors and students are available. Given that in this study we focused on the aspect of teaching performance, the sections 6, 7 and 8 of the survey were not taken into account.

An initial inspection of the data was carried out and it was observed that the lowest categories in the two scales had few answers, so those categories were collapsed and the items were recoded into four ordinal response categories.

According to the proposed model, as the effect of the course is one facet, we debugged the data to obtain professors who were evaluated in at least two courses. We obtained 203 professors with such condition. A sample of 50 professors was selected to run the example. In that sample, a total of 2505 student responses were reported, corresponding to 94 courses.

6.2. Unidimensionality analysis

According to presented in section 5.1, we verify unidimensionality with the three tools: principal component analysis (Jollife 2002) and reliability analysis

through Cronbach's alpha coefficient (Cronbach 1951) by using the Cronbach-Mesbah curve (Cameletti & Caviezel 2012).

The left panel in figure 2 shows the bar chart of the percentage of variance explained by the eigenvalues, and the right panel presents the Cronbach-Mesbah curve.

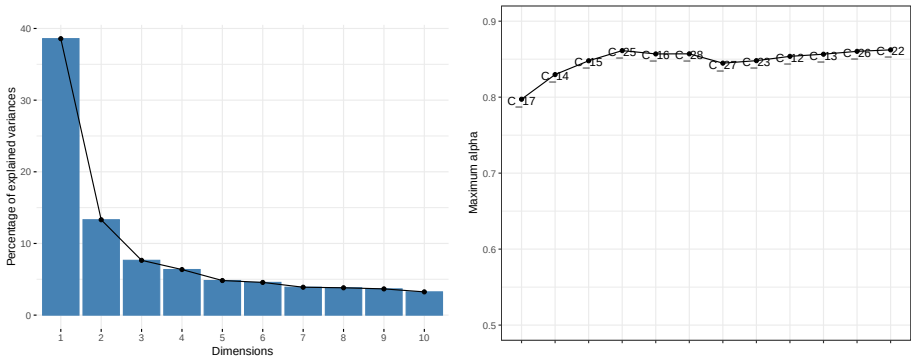

FIGURE 2: Verification of unidimensionality at the Universidad Nacional survey data

The figure in the left panel suggests that the data are unidimensional since the percentage of variance explained by the first own value is substantially greater than the percentage explained by the second value. On the other hand, in the right panel, the Cronbach-Mesbah curve shows the Cronbach coefficient after each item that maximizes the reliability of the data is removed.

In accordance with Cameletti & Caviezel (2012), if the data is unidimensional, then the curve is monotonously increasing. However, although the curve does not have strictly that behavior, it is observed that the change in reliability when extracting two items is negligible in terms of the scale. It can be verified that the estimated values of Cronbach's Alpha in the curve are between 0.79 and 0.87. Given the above, it is concluded *carefully* that the instrument complies with the one-dimensional assumption. In addition, the estimated value of Cronbach's Alpha for the sample is 0.86, therefore it is presumed that the instrument is measuring a single construct.

6.3. Parameter Estimation

The proposed model to fit the data was run in Stan (Stan Development Team 2020*c*, Stan Development Team 2020*b*), through its R interface named Rstan (Stan Development Team 2020*a*). Stan is a probabilistic programming language currently used to run very complex Bayesian codes at high performance and is a free software which implements Hamiltonian MCMC samplers (Neal 2011) and the No-U-turn sampler (Hoffman & Gelman 2014). For details on the code refer to the Appendix B. This section specifies the most relevant results related to the specification of the parameters.

6.3.1. Severity parameters

The inclusion of a severity parameter in the model allows capturing the variability associated with the answers given by the students, determining for each of them the degree of severity or indulgence they perceive the performance of each professor. Negative values of the parameter imply that the evaluator has a tendency to give higher grades than the average, while positive values imply that the evaluator has tendency to give low grades.

Figure 3 presents the distribution of the severity parameter γ for the 2505 students and the credibility bands ordered for some of them.

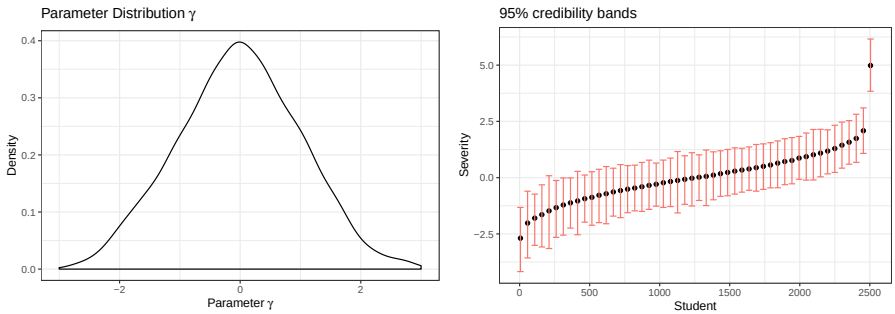

FIGURE 3: Results of the severity parameter estimate γ

The distribution of the severity parameters is approximately symmetrical, the model suggests that there are forgiving students in the sample with a high degree of severity in the same proportion.

Additionally, figure 4 presents box plots showing the distributions of the severity parameters estimated according to the gender of the students and to the type of program they have with the university, that is, if they are part of an undergraduate or postgraduate program.

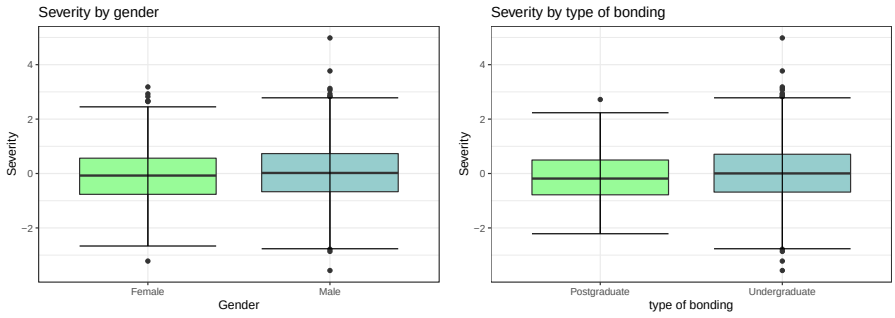

FIGURE 4: Results of severity parameter according to gender and type of program

From the above, visually it is observed that there are no substantial differences between the grades given by female students with those of male. However, the

average severity of women is -0.072 while that of men is 0.027, which suggests that men are slightly less forgiving when rating teaching performance.

To illustrate if there are differences between these averages, a mean difference test is performed and a p-value of 0.023 is obtained, which indicates that if there is a difference between the average severity of women and men, however, it is important to note that the sample sizes for this test are large (853 women and 1652 men) and therefore this test will be significant in favor of the difference.

For the type of program, differences are observed between the two distributions, however, for this case, the distribution of the severities of the undergraduate students has greater variability. The average severity of graduate students is -0.164 while that of undergraduate students is 0. This suggests that graduate students are on average more forgiving than undergraduate students. Likewise, when performing a test of difference in measures, a p-value of 0.105 is obtained, which indicates that there are differences between the average severity of undergraduate and graduate students.

6.3.2. Course parameters

The course parameter improves the adjustment of the estimation of teaching performance by controlling the effect of the characteristics of the course. This parameter is interpreted as a general difficulty of the course, therefore, estimated negative values of the parameter indicate that the course has a low perception of difficulty, while positive values indicate that the course has a high perception of difficulty.

Figure 5 shows the distribution of the parameter for the 97 courses and its credibility bands. The distribution of the course parameters η has an approximately symmetrical distribution, the estimated values are in a range of -1,087 to 0.997, with an estimated average value of -0.056 and a deviation value of 0.473 . The distribution presents a low variability, which indicates that the estimated difficulty of the courses is concentrated around 0. Due to the symmetry shown in the graph, the model suggests that in the sample there are courses with a perception of low and high difficulty approximately in the same proportion.

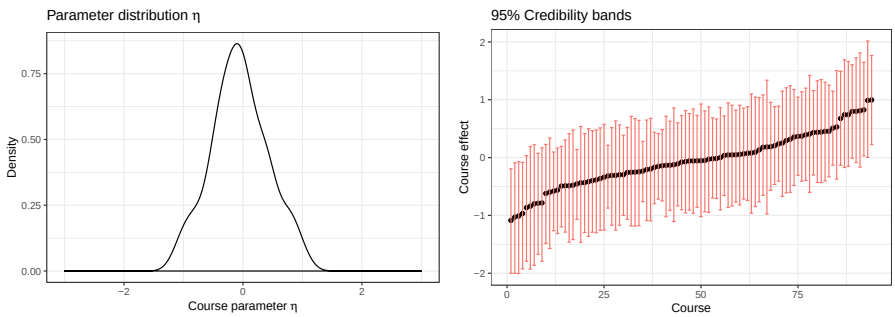

FIGURE 5: Course parameters η results

6.3.3. Estimation of teaching performance and item parameters

The parameter θ related to teacher performance is one of the most important in the structure of the model, this represents the perception of teacher performance in the classroom. Negative values of the parameter indicate a low perception of teacher performance according to the criteria of the students, while positive values indicate a high perception of performance.

The figure 6 presents in the left panel the distribution of the parameter and in the right panel presents the credibility bands of 95% ordered for each of the 50 professors in the sample.

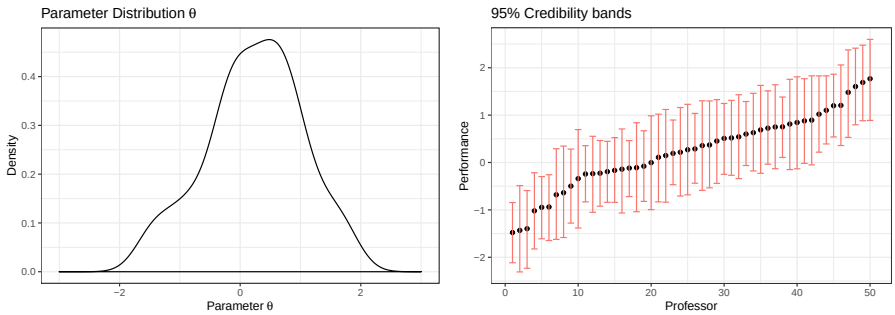

FIGURE 6: Results teacher performance θ

The posterior distribution has an average value of 0.235 and a standard deviation of 0.796, evidently it is a positive asymmetric distribution. In general, the perception of teacher performance in the sample is good, however, there is a small set of teachers with a very low perception of performance compared to the average.

Figure 7 shows the distribution of skills by teacher gender. The distribution for the female gender is more variable than for the male gender. The estimated average performance in teachers is perceived higher than in teachers, with values of 0.501 and 0.105, respectively.

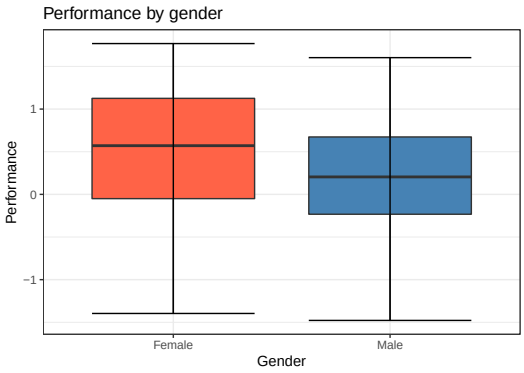

FIGURE 7: Distribución de habilidades estimadas por género

A mean difference test between the average performance of teachers by gender was made, a p value of 0.094 was obtained. This value indicates that there is a difference between the estimated average values. The previous results suggest that in the Faculty of Sciences there are significant differences between the perception of the teachers' teaching performance in favor of the females.

The figure 8 shows the behavior of the items in the test with respect to the parameter β. Each line represents an item, and the red points are those estimated cutoff points β_k. Each item has 3 values of β_k, naturally because each one has 4 response options.

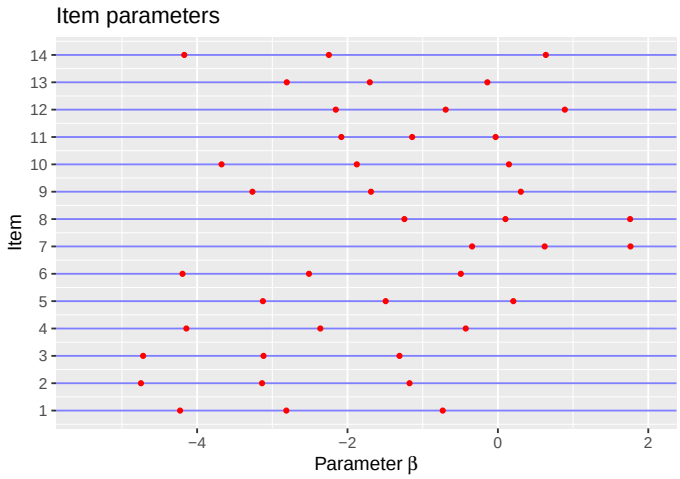

FIGURE 8: Items parameters

It is observed that the distribution of the cutoff points β_k of the items are concentrated in values between -5 and 2, that is, values skewed to the left. Taking this into account, only two items have their highest value of β_k near to 2. This is an indicator that the test in general has a slightly skewed behavior to the left, that is, that the items are perceived with a low difficulty for measuring the perception of teaching performance.

Items 2 and 3 have the lowest difficulty category, in these two, students tend to give teachers very high marks. In contrast, items 7 and 8 presents the highest difficulty category of the questionnaire since the students gave homogeneously lower grades. The estimated values of the item parameters can be seen in the Appendix C.

6.4. Model fit

Section 5.2 describes a statistic for evaluating the goodness of fit of the model. In this section, the adjustment for the estimated parameters of ability θ and the items β_k is reviewed, as well as the obtaining of the goodness of fit statistics at the global level of the model.

The value of the global goodness-of-fit statistic in the proposed model was 0.41. The average of the goodness of fit statistics for the items in the proposed model was 0.43 and the average of the goodness-of-fit statistics values for the skills of those evaluated was 0.55. According to the established criteria, these values indicate good fit of the models.

6.5. Selection criteria models

In this section, we made a comparison with a model that has fewer parameters. For practical purposes, the comparison model is called *model zero*.

Technically, the zero model in its mathematical form does not include the additional parameter η, that is, this multi-faceted model does not recognize the variability associated with the course taught by the teacher. The conditional probability model is defined as:

$$Pr\left[Y_{ijs} = k|\theta_i, \gamma_{is}, \beta_j\right] = \text{logit}^{-1}(\theta_i - \gamma_{is} - \beta_{j(k-1)}) - \text{logit}^{-1}(\theta_i - \gamma_{is} - \beta_{jk}) \tag{12}$$

where Y_{ijs} is the random variable that represents the score assigned by the student s to the teacher i in the item j, θ_{ic} is the latent trace or performance of the teacher i, β_j the difficulty of the item j and γ_{is} the severity of the student s to the teacher i.

Given the previous expression, and making a revision of the expression (1) where the conditional probability of the proposed model is defined, it can be deduced that the proposed model is a generalization of the zero model because the objective of the inclusion of the parameter η is to capture additional variability that should not influence the estimation of the skill parameter θ.

For each of the models, the estimation of the statistics of the selection criteria is performed. The table 1 presents the results obtained for the WAIC and LOO criteria detailed in section 5.3. These criteria are generally more robust in hierarchical models or models with random effects.

Model	**WAIC**	**LOO**
Zero model	64808,8	64827,7
Proposed model	64788,4	64806,6

TABLE 1: Criteria for model selection

The estimated values for the two models are close, however, it is important to mention that the gain obtained by including the additional parameter in terms of isolating the variability associated with the course is important. Statistical values are lower for both criteria for the proposed model. Given this, it is inferred that according to the selection criteria the model to be selected is the proposed model because it presents lower values.

7. Conclusions

The problem of estimating teaching performance from evaluations of student perception was addressed. The evidence found showed that, in general, the institutions make an incorrect use of the results of the surveys of perception of teaching performance, this, because they carry out averages of ordinal scales assigning a numerical value without any theoretical support.

The main objective of this document was to give a statistical approach to the use of evaluations of perception of teacher performance from the use of TRI models, in order to obtain estimates of latent traces for both the evaluated As for the evaluator, this taking into account that the evaluation involves aspects that can influence which cannot be controlled.

The evidence shows that the implementation of a statistical model is important in two ways: 1) It allows making inferences about the estimated parameters associated with the model, in addition to having an associated error measurement, and 2) It captures variability associated with the characteristics of the evaluation that are not they are taken into account when other procedures are carried out, that is, the course parameter is important to isolate the effect of the course that the students take and that they are indirectly measuring when evaluating the teacher.

The problem was addressed with data from a survey of students on the perception of teaching performance at the Faculty of Sciences of the National University of Colombia. The final instrument after the review had 14 items focused on topics that inquired about the perception of teacher performance in the classroom.

The application had two orientations, the first focused on the fit of the model and the analysis of the results, and the second focused on the fit of the model and in comparing the proposed model with a model called zero model that does not considered in its mathematical expression the parameter associated with the course η. The measurement instrument complies with the assumption of one-dimensionality, that is, all the items are measuring a construct, which for the case is it can be called *Perception of teacher performance*.

It is observed that the severity parameter of the model it has a symmetrical behavior in its distribution, which indicates that there is an appropriate range of severities for the measurement of teachers' skills. The importance of including this parameter is that there are characteristics associated with students that influence the grade they give to students, such as, for example, gender or the type of relationship they have with the institution, in this case, it was obtained that in general, female students are more lenient, as well as, students associated with graduate programs are also more forgiving than those enrolled in undergraduate programs.

About the course parameter, a symmetrical distribution with low variability is also observed, that is, the majority of parameters are around the value 0. The inclusion of this parameter in the model is of great importance because this parameter captures the effect associated with the student's perception of the course and isolates this effect from the estimation of the teacher's ability parameter.

On the other hand, the model was evaluated in two ways: First, it was observed that the model fit statistic is close to 0.5 (with a value of 0.41), which indicates that the model fits the data well. Second, a review of model selection criteria compared to a model with a fewer parameters model, it was obtained that the proposed model has lower values in the WAIC and LOO statistics, which is an indicator that the inclusion of the additional parameter η related to the effect of the course on the model fits best and is appropriate for the data.

Taking into account the above, it is concluded that the implementation of the methodology proposed in this document is very useful, because it statistically addresses a problem that many institutions face when evaluating their teachers, this in the sense of improving assessment instruments and improving the teacher evaluation process itself.

Acknowledgements

This study was supported by the Universidad Nacional de Colombia, grant number 36008. It is part of a project focused on the *Development and implementation of a multifaceted item response model for analysis of professor evaluation at the Universidad Nacional de Colombia.* This is a work of the SICS Research Group of the Universidad Nacional de Colombia.

References

Abrami, P. C., Perry, R. P. & Leventhal, L. (1982), 'The relationship between student personality characteristics, teacher ratings, and student achievement.', *Journal of Educational Psychology* **74**(1), 111.

Ariyo, O., Quintero, A., Muñoz, J., Verbeke, G. & Lesaffre, E. (2019), 'Bayesian model selection in linear mixed models for longitudinal data', *Journal of Applied Statistics* pp. 1–24.

Baker, F. B. & Kim, S. H. (2004), *Item Response Theory*, 2nd edn, Marcel Decker Inc.

Barkaoui, K. (2014), *Multifaceted Rasch analysis for test evaluation*, Chichester, UK: Wiley, pp. 1301–1322.

Bartholomew, D., Knott, M. & Moustaki, I. (2011), *Latent Variable Models and Factor Analysis. A Unified Approach*, third edn, Wiley.

Basow, S. A. & Silberg, N. T. (1987), 'Student evaluations of college professors: Are female and male professors rated differently?', *Journal of educational psychology* **79**(3), 308.

Becker, W. E. & Watts, M. (1999), 'How departments of economics evaluate teaching', *American Economic Review* **89**(2), 344–349.

Bélanger, C. H. & Longden, B. (2009), 'The effective teacher's characteristics as perceived by students', *Tertiary Education and Management* **15**(4), 323–340.

Birnbaum, A. (1968), *Statistical Theories of mental test Scores*, Reading, MA: Addison Wesley, chapter Trait models and their use in infering an examinee's ability.

Bock, R. D. (1997), 'A brief history of item response theory', *Educational Measurement: Issues and Practice* **16**(4), 21–32.

Box, G. E. (1980), 'Sampling and bayes' inference in scientific modelling and robustness', *Journal of the Royal Statistical Society: Series A (General)* **143**(4), 383–404.

Braga, M., Paccagnella, M. & Pellizzari, M. (2014), 'Evaluating students' evaluations of professors', *Economics of Education Review* **41**, 71–88.

Cameletti, M. & Caviezel, V. (2012), 'The cronbach-mesbah curve for assessing the unidimensionality of an item set: The r package cmc'.

Centra, J. A. (1993), *Reflective Faculty Evaluation: Enhancing Teaching and Determining Faculty Effectiveness. The Jossey-Bass Higher and Adult Education Series.*, ERIC.

Centra, J. A. & Creech, F. R. (1976), The relationship between student teachers and course characteristics and student ratings of teacher effectieness, *in* 'Project Report', Princeton, NJ, Educational Testing Service, pp. 76–1.

Cohen, P. A. (1981), 'Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies', *Review of educational Research* **51**(3), 281–309.

Cordoba, K. (2020), Un modelo tri de múltiples facetas para la evaluación del desempeño docente en el aula, Master's thesis, Universidad Nacional de Colombia.

Cronbach, L. J. (1951), 'Coefficient alpha and the internal structure of tests', *Psychometrika* **16**, 297–334.

Eckes, T. (2015), *Introduction to Many-Facet Rash Measurement. Analyzing and Evaluating Rater-Mediated Assesments*, second edn, Peter Lang Edition.

Engelhard, G. (2002), *Monitoring raters in performance assessment*, Mahwah, NJ: Erlbaum., pp. 261–287.

Engelhard, G. (2013), *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*, New York, NY: Routledge.

Feldman, K. A. (1977), 'Consistency and variability among college students in rating their teachers and courses: A review and analysis', *Research in Higher Education* **6**(3), 223–274.

Feldman, K. A. (1978), 'Course characteristics and college students' ratings of their teachers: What we know and what we don't', *Research in Higher Education* **9**(3), 199–242.

Feldman, K. A. (1979), 'The significance of circumstances for college students' ratings of their teachers and courses', *Research in Higher Education* **10**(2), 149–172.

Feldman, K. A. (1983), 'Seniority and experience of college teachers as related to evaluations they receive from students', *Research in Higher Education* **18**(1), 3–124.

Feldman, K. A. (1987), 'Research productivity and scholarly accomplishment of college teachers as related to their instructional effectiveness: A review and exploration', *Research in higher education* **26**(3), 227–298.

Feldman, K. A. (1989), 'The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies', *Research in Higher education* **30**(6), 583–645.

Gelfand, A. E., Dey, D. K. & Chang, H. (1992), Model determination using predictive distributions with implementation via sampling-based methods, Technical report, Stanford University CA Department of statistics.

Gelman, A., Hwang, J. & Vehtari, A. (2014), 'Understanding predictive information criteria for bayesian models', *Statistics and computing* **24**(6), 997–1016.

Gelman, A., Meng, X.-L. & Stern, H. (1996), 'Posterior predictive assessment of model fitness via realized discrepancies', *Statistica sinica* pp. 733–760.

Hoffman, M. D. & Gelman, A. (2014), 'The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo.', *Journal of Machine Learning Research* **15**(1), 1593–1623.

Jollife, I. (2002), *Principal Component Analysis*, 2nd edn, Springer.

Koushki, P. A. & Kunh, H. A. J. (1982), 'How realiable are student evaluations of teachers?', *Engineering Education* **72**, 362–367.

Linacre, J. M. (1989), *Many-facet Rasch measurement*, Chicago: MESA Press.

Lord, F. & Novick, M. (2013), *Statistical Theories of Mental Test Scores*, Addison-Wesley Publishing Company.

Luo, Y. & Jiao, H. (2018), 'Using the stan program for bayesian item response theory', *Educational and psychological measurement* **78**(3), 384–408.

Marsh, H. W. (1987), 'Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research', *International journal of educational research* **11**(3), 253–388.

Marsh, H. W. (2007), Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness, *in* 'The scholarship of teaching and learning in higher education: An evidence-based perspective', Springer, pp. 319–383.

Martin, E. (1984), 'Power and authority in the classroom: Sexist stereotypes in teaching evaluations', *Signs: Journal of Women in Culture and Society* **9**(3), 482–492.

Murray, H. G. (2005), Student evaluation of teaching: Has it made a difference, *in* 'Annual Meeting of the Society for Teaching and Learning in Higher Education. Charlottetown, Prince Edward Island'.

Neal, R. (2011), *MCMC using Hamiltonian dynamics in Handbook of Markov Chain Monte Carlo*, New York, NY: CRC Press., pp. 113–162.

Perry, R. P., Niemi, R. R. & Jones, K. (1974), 'Effect of prior teaching evaluations and lecture presentation on ratings of teaching performance.', *Journal of Educational Psychology* **66**(6), 851.

Small, A. C., Hollenbeck, A. R. & Haley, R. L. (1982), 'The effect of emotional state on student ratings of instructors', *Teaching of Psychology* **9**(4), 205–211.

Spencer, P. A. & Flyr, M. L. (1992), 'The formal evaluation as an impetus to classroom change: Myth or reality?.'.

Stan Development Team (2020*a*), 'RStan: the R interface to Stan'. R package version 2.19.3.
*http://mc-stan.org/

Stan Development Team (2020*b*), 'Stan language reference manual'. Version 2.22.
*http://mc-stan.org

Stan Development Team (2020*c*), 'Stan user's guide'. Version 2.22.
*http://mc-stan.org

Stark, P. & Freishtat, R. (2014), 'An evaluation of course evaluations', *ScienceOpen Research* .

Uttl, B., Eche, A., Fast, O., Mathison, B., Valladares Montemayor, H. & Raab, V. (2012), 'Student evaluation of instruction/teaching (sei/set) review', *Calgary, AB, Canada: Mount Royal Faculty Association Retrieved from: http://mrfa. net/files/MRFA_ SEI_ Review_ v6. pdf* .

Uttl, B., White, C. A. & Gonzalez, D. W. (2017), 'Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related', *Studies in Educational Evaluation* **54**, 22–42.

Vehtari, A., Gelman, A. & Gabry, J. (2017), 'Practical bayesian model evaluation using leave-one-out cross-validation and waic', *Statistics and computing* **27**(5), 1413–1432.

Wachtel, H. K. (1998), 'Student evaluation of college teaching effectiveness: A brief review', *Assessment & Evaluation in Higher Education* **23**(2), 191–212.

Watanabe, S. (2010), 'Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory', *Journal of Machine Learning Research* **11**(Dec), 3571–3594.

Appendix A. Questionnaire used to Measure Professor Performance

The original scale was 1 2 3 4 5. Categories 1 and 2 were merged, so, in the end, there were 4 categories for each question.

1. Does the professor attend classes regularly and punctually?
2. Does the professor respect the agreed dates for academic activities, including evaluations and delivery of results?
3. Does the professor prepare each of the sessions of the course beforehand?
4. Is the professor accessible and willing to provide academic help?
5. Does the professor encourage group work, recognizing student successes and achievements during learning activities?
6. Does the professor demonstrate commitment and enthusiasm in their teaching activities?
7. Do you consider the content of the subject to be clear and specific?
8. Does the professor include learning experiences in places other than the classroom (e.g., workshops, laboratories, companies, the community, etc.)?
9. Does the professor organize activities that allow students to exercise oral and written expression?
10. Does the professor develop the content of the class in an orderly and understandable manner?
11. Does the professor promote self-study and research?
12. Does the professor use technology (e.g., computer, video beam, digital platforms, e-mail, etc.) as a means to facilitate student learning?
13. Does the professor promote the use of various digital tools to manage (collect, process, evaluate and use) information?
14. Does the professor promote the safe, legal and ethical use of digital information?
15. In general, the professor's performance was?

Appendix B. Stan Code for Estimating Professor Performance

A Stan program is organized in a sequence of blocks, whose contents are declarations of variables. Each component of the model for the code in Stan is described below (Luo & Jiao 2018).

Data block

Data block is the first component in a Stan code. The required data for model estimation is specified there. In this work, information related to the number of responses, professors, students, items and categories for the item, the number of courses and the responses must be specified in this section. The code used for the data block is listed below:

```
data{
    int<lower=1>  N;                      // # of responses
    int<lower=10> N_prof;                 // # of professors
    int<lower=10>  N_stud;                //  # of students
    int<lower=2>  N_item;                 // # of items
    int<lower=2,upper=5> N_cat;           // # of categories by item
    int<lower=10> N_sub;                   // # of courses
    int<lower=1,upper=N_cat> y[N];        // y[n], n-th response
    int<lower=1,upper=N_prof> professor[N];// Response professor n
    int<lower=1,upper=N_item> item[N];     // Response item n
    int<lower=1,upper=N_stud> student[N];  // Response student n
    int<lower=1,upper=N_sub> subject[N];  // Response course n
    }
```

Parameter block

In this block the model parameters are specified. For the context of multifaceted TRI models,item, performance, course and severity parameters must be specified. On the other hand, it is also necessary to specify the hyper-parameters associated to the variances in the model. The variables declared in this block correspond to the variables to be sampled. The following code was used in this work.

```
parameters{
   vector[N_prof] theta;         // Latent trait of the professor
   vector[N_stud] gamma;         // Student severity
   vector[N_sub] eta;            // Course parameter
   ordered[N_cat-1] beta[N_item];// Item parameters
   real<lower=0> sigma_eta;      // Deviation of the parameter eta
   real<lower=0> sigma_gamma;    // Deviation of the parameter gamma
   real<lower=0> sigma_beta;     // Deviation of the parameters beta
}
```

Model block

In this block the model is implemented. All prior distributions and the likelihood are declared in this section. The model block of our model is as follows.

```
23  model{
24    theta ~ normal(0,1);               // Prior of latent traits
25    eta ~ normal(0,sigma_eta);         // Prior of course parameter
26    gamma ~ normal(0,sigma_gamma);     // Prior of severity parameter
27    for(j in 1:N_item){
28      beta[j] ~ normal(0,sigma_beta);// Prior of item parameters
29    }
30    sigma_beta ~ cauchy(0,2);   // Hyer-prior for sigma_beta
31    sigma_gamma ~ cauchy(0,2);  // Hyper-prior for sigma_gamma
32    sigma_eta ~ cauchy(0,2);    // Hyper-prior for sigma_eta
33    for(n in 1:N){
34      y[n] ~ ordered_logistic(theta[professor[n]]-eta[subject[n]]
35             -gamma[student[n]], beta[item[n]]); //likelihood
36    }
37   }
```

Generated quantities block

This block does not affect the values of the sampled parameters. If a quantity does not play a role in the model, it must be defined in the block of generated quantities. In this case the associated log-likelihood is calculated. Our code is as follows.

```
38  generated quantities{
39    vector[N] log_lik;  // Log-likelihood vector
40      for(n in 1:N){
41        log_lik[n] = ordered_logistic_lpmf(y[n]|theta[professor[n]]
-eta[subject[n]]-gamma[student[n]], beta[item[n]]);  // log-likelihood
42    }
43  }
```

Appendix C. Estimated Item Parameters in Professor Performance Measurement

Table 2 shows the results obtained from the estimation process of the item transformed parameters, using Stan The $\hat{R}$ statistics was omitted, because all its values are very close to 1.

TABLE 2: Estimate β-parameters of professor performance, computed by Stan. Column Rhat ($\widehat{R}$) is omited. All values are very close to 1.00. Column *n eff* is the effective sample size.

Parameter	**Mean**	**Std. Dev.**	**q2.5**	**q25**	**q50**	**q75**	**q97.5**	**n_eff**
$\beta_{1,1}$	-4,23	0,21	-4,62	-4,37	-4,23	-4,09	-3,83	657,18
$\beta_{1,2}$	-2,81	0,18	-3,16	-2,94	-2,81	-2,69	-2,46	526,27
$\beta_{1,3}$	-0,73	0,17	-1,07	-0,85	-0,73	-0,62	-0,40	442,49
$\beta_{2,1}$	-4,75	0,22	-5,19	-4,90	-4,74	-4,59	-4,32	720,17
$\beta_{2,2}$	-3,14	0,18	-3,50	-3,26	-3,14	-3,01	-2,78	519,68
$\beta_{2,3}$	-1,17	0,17	-1,51	-1,29	-1,18	-1,06	-0,84	465,72
$\beta_{3,1}$	-4,72	0,22	-5,15	-4,87	-4,72	-4,57	-4,29	715,35
$\beta_{3,2}$	-3,12	0,18	-3,48	-3,25	-3,12	-2,99	-2,76	520,17
$\beta_{3,3}$	-1,31	0,17	-1,64	-1,43	-1,31	-1,19	-0,98	474,06
$\beta_{4,1}$	-4,14	0,20	-4,54	-4,28	-4,14	-4,00	-3,76	538,25
$\beta_{4,2}$	-2,36	0,18	-2,70	-2,48	-2,36	-2,24	-2,02	477,77
$\beta_{4,3}$	-0,43	0,17	-0,75	-0,55	-0,43	-0,31	-0,10	452,42
$\beta_{5,1}$	-3,13	0,18	-3,48	-3,25	-3,13	-3,00	-2,76	477,28
$\beta_{5,2}$	-1,49	0,17	-1,83	-1,61	-1,49	-1,37	-1,16	447,19
$\beta_{5,3}$	0,20	0,17	-0,12	0,09	0,20	0,32	0,53	435,69
$\beta_{6,1}$	-4,19	0,20	-4,58	-4,33	-4,19	-4,05	-3,80	599,74
$\beta_{6,2}$	-2,51	0,18	-2,86	-2,64	-2,51	-2,39	-2,17	452,54
$\beta_{6,3}$	-0,49	0,17	-0,82	-0,61	-0,50	-0,37	-0,17	435,13
$\beta_{7,1}$	-0,34	0,17	-0,67	-0,46	-0,34	-0,23	-0,01	450,46
$\beta_{7,2}$	0,62	0,17	0,29	0,51	0,62	0,74	0,95	444,26
$\beta_{7,3}$	1,76	0,17	1,43	1,64	1,76	1,88	2,09	457,30
$\beta_{8,1}$	-1,24	0,17	-1,57	-1,36	-1,24	-1,13	-0,91	456,98
$\beta_{8,2}$	0,10	0,17	-0,23	-0,02	0,10	0,22	0,43	443,21
$\beta_{8,3}$	1,76	0,17	1,43	1,64	1,76	1,87	2,09	460,74
$\beta_{9,1}$	-3,26	0,19	-3,63	-3,39	-3,26	-3,14	-2,90	512,49
$\beta_{9,2}$	-1,69	0,17	-2,03	-1,81	-1,69	-1,57	-1,35	432,42
$\beta_{9,3}$	0,31	0,17	-0,02	0,19	0,30	0,42	0,63	434,23
$\beta_{10,1}$	-3,67	0,19	-4,04	-3,81	-3,67	-3,54	-3,30	537,18
$\beta_{10,2}$	-1,88	0,17	-2,21	-2,00	-1,88	-1,76	-1,54	426,88
$\beta_{10,3}$	0,15	0,17	-0,18	0,03	0,15	0,26	0,48	424,03
$\beta_{11,1}$	-2,08	0,17	-2,42	-2,20	-2,08	-1,96	-1,75	469,91
$\beta_{11,2}$	-1,14	0,17	-1,47	-1,26	-1,14	-1,02	-0,82	437,02
$\beta_{11,3}$	-0,03	0,17	-0,36	-0,15	-0,03	0,08	0,30	447,72
$\beta_{12,1}$	-2,16	0,17	-2,50	-2,28	-2,16	-2,03	-1,82	469,99
$\beta_{12,2}$	-0,69	0,17	-1,02	-0,81	-0,69	-0,58	-0,36	452,13
$\beta_{12,3}$	0,89	0,17	0,57	0,77	0,89	1,01	1,22	465,29
$\beta_{13,1}$	-2,81	0,18	-3,15	-2,93	-2,81	-2,68	-2,47	474,87
$\beta_{13,2}$	-1,70	0,17	-2,04	-1,82	-1,70	-1,59	-1,37	446,79
$\beta_{13,3}$	-0,14	0,17	-0,47	-0,26	-0,14	-0,02	0,19	453,58
$\beta_{14,1}$	-4,17	0,20	-4,56	-4,31	-4,17	-4,03	-3,78	573,60
$\beta_{14,2}$	-2,25	0,18	-2,59	-2,37	-2,25	-2,13	-1,90	449,51
$\beta_{14,3}$	0,64	0,17	0,32	0,52	0,64	0,76	0,96	439,46