Assessment and Evaluation in Higher Education manuscript No.
(will be inserted by the editor)

Bayesian Multi-faceted models for measuring classroom professor performance and subject relevance

Alvaro Montenegro, Harvey Rosas, Campo Elías Pardo, Eduardo Jorquera, Cristian Millán, Jhonatan Medina

Received: date / Accepted: date

Abstract There is great debate about the validity of classical measurements of a professors performance from student evaluations of teaching (SET). Despite criticism, many universities use SET to evaluate their faculty's teaching effectiveness. A significant problem with those measures is that, in general, they are invalid from a mathematical standpoint; thus, analyses based on the results of those measures are questionable. To solve this issue, this study proposed that Bayesian multi-faceted models be applied to SET data sets. The models, including professor parameters, questionnaire parameters and student severity parameters, were applied to a data set from a course evaluation at the Universidad Nacional de Colombia. The data included items about professor performance, as well as course related questions. A strategic diagram illustrating both the professors performance and subjects relevance is introduced as an analysis tool and aid for decision-making related to professor performance.

Keywords Multi-faceted models, professor evaluation, student surveys

Alvaro Montenegro, corresponding author
Departamento de Estadística, Universidad Nacional de Colombia,
E-mail: ammontenegrod@unal.edu.co,
Harvey Rosas
Departamento de Estadística, Universidad de Valparaíso,
E-mail: hjrosasq@uv.cl,
Campo Elías Pardo
Departamento de Estadística, Universidad Nacional de Colombia,
E-mail: cepardot@unal.edu.co
Eduardo Jorquera
Departamento de Estadística, Universidad de Valparaíso,
E-mail: eduardoejorqueras@posgrado.uv.cl,
Cristian Millán
Departamento de Estadística, Universidad Nacional de Colombia,
E-mail: ccmillana@unal.edu.co,
Jhonatan Medina
Departamento de Estadística, Universidad Nacional de Colombia,
E-mail: jmedinau@unal.edu.co.

1 Introduction

Student evaluations of teaching (SET) have been used and debated for a long time (Stark and Freishtat, 2014). SET are often used to aid in hiring and promotion decision-making processes, (Becker and Watts, 1999). However, there have been many criticisms about the results. In fact, the majority of researchers agree that SET ratings and student learning are not related (Uttl et al., 2017). Apparently, there is strong evidence that student responses to questions of effectiveness are not, in fact, an accurate measure teaching effectiveness (Stark and Freishtat, 2014). On the other hand, when comparing SET scores with grades students expected to receive, the reported results often conflict. In a radical affirmation, Krautmann and Sander (1999) suggest that instructors could buy better evaluations through more lenient grading. In general, some studies suggest that there is a positive relationship between SET and expected grades (Aigner and Thum, 1986; Ditts, 1980; Mehdizadeh, 1990); however, other studies find no such relationships (Seiver, 1983; Decanio, 1986).

However, in spite of such criticism, a lot of universities use SET to evaluate their facultys teaching effectiveness (Wachtel, 1998; Uttl et al., 2017). Typically, SET are conducted within the last two weeks of a course. Students are presented with forms that ask them to rate their perceptions about instructors and courses, often on a 5-point Likert scale.

The validity of anonymous SET is based on the assumption that students can observe the professors performance in the classroom and will report it truthfully when asked. However, some studies suggest that SET seems to be influenced by a variety of factors unrelated to teaching performance, including the gender, ethnicity, and even the attractiveness of the professor (Stark and Freishtat, 2014). On the other hand, the students' objectives might be different from those of the administration (Braga et al., 2014). For example, students may be more interested and motivated in core courses than in complementary ones, which is often reflected in how they evaluate their professors. As a result, those in core courses might end up with better evaluations.

The discussion in this paper focuses on the validity of the current measures obtained from SET data and proposes a special kind of Bayesian model to apply to SET data. However, the topic of teaching effectiveness is not specifically addressed within this paper.

When experts report the results of their research about SET, the documentation often includes all the different variables introduced in their models. However, information about the algorithm used to compute the reported *professors' performance* is often omitted. In practice, people involved in the evaluation of professors routinely compare professors average scores to departmental averages, comparisons which are statistically questionable (Stark and Freishtat, 2014). SET scores are ordinal categorical variables; they have a natural order, but the distance concept makes no sense. Although the Likert scale has the nomenclature 1, 2, and so on, it is not numeric. Thus, the difference between 3 and 4 is not comparable with the difference between 4 and 5. Furthermore, labels A, B,... could be used in an equivalent way. As a result, averages based on this kind of data make no sense.

An additional problem is associated with the precise algorithm used to compute each performance score. It is not clear how the missing values are treated nor how the score is computed. The score may be a global average based on all the responses of the students, or it may be computed in two stages. First, an average is calculated for each student, and second, a harmonic mean from these averages may be obtained. On the other hand, the questions may have a different number of categories than the potential responses; for example, some of them could be binary requiring only a yes or no answer, while others require responses on a scale from 1 to 5.

Many people, especially among faculty administration, feel at ease with SET because numerical values are obtained from computing the scores. However, in general, any type of arithmetic procedure with ordinal variables is considered mathematically invalid. An exception is a case in which all variables are binary, as in dichotomized educational tests; in this case, the sum corresponds to the number of correct responses.

Consequently, statistical models where these types of averages are treated as continuous variables are questionable, as are the studies based on such models. Obviously, such studies may reflect the truth; however, if a model has a questionable construction, conclusions based on it are not consistent.

On the other hand, ordinal variables can be legally included in some statistical models like logit or multinomial logistic regression. In this paper, new models to be applied to SET data are introduced, which were used them for measuring professors performance, as well as the course relevance. The proposed models were inspired by the multi-faceted Rach models (Eckes, 2015). In this study, Bayesian models were used as an alternative to Rach models, given the lack of previous work using these techniques to analyze SET data.

The proposed models allow for the design of measures that are valid from both the mathematical and statistical point of view. These models ensure the quality of the performance measurements with the available data. Explanatory variables may be introduced in the models to avoid the biases mentioned.

2 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, 2013), creative performance assessment (Pin et al., 2012), evaluation of student and teacher presentations, (Peters et al., 2010), analysis of open-ended questions (Guler, 2014), 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 common practice, raters are required to score examinees using a specific rubric. Nevertheless, raters may influence examinees scores in a number of ways. Wolfe and Chiu (1997) discuss how rater influence can be detected with a multi-faceted rating scale model.

On the other hand, raters introduce variability into the scores awarded 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

$$\begin{aligned}
p_{nljk} &= \text{ prob. of examinee } n \text{ reciving a rating } k \text{ from rater } j \text{ on task } l,\\
p_{nljk-1} &= \text{ prob. of examinee } n \text{ reciving a rating } k-1 \text{ from rater } j \text{ on task } l,\\
\theta_n &= \text{ ability of examine } n,\\
\delta_l &= \text{ difficulty of task } l,\\
\alpha_j &= \text{ severity of rater } j,\\
\tau_k &= \text{ difficulty of receiving a rating of } k \text{ relative to } k-1.
\end{aligned}$$

3 Ordered Logistic Distribution

In this section, the ordered logistic distribution is introduced. The definitions of the Bayesian multi-faceted models introduced in the next section are based on this 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. The multi-faceted models are item response models. Thus, 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 hand, 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.

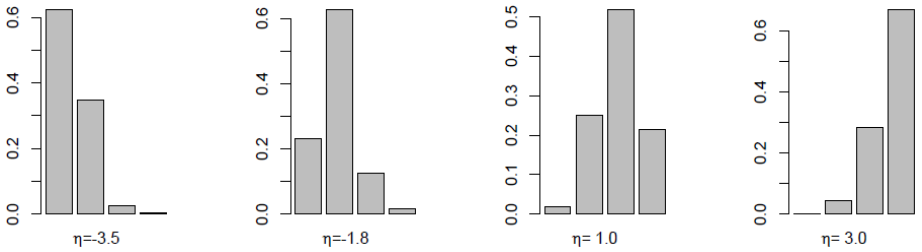

Fig. 1 Probabilistic mass function of the Ordered Logistic distribution for different values of η, with $\boldsymbol{\beta} = (-3.0, 0.0, 2.3)^t, K = 4$.

4 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.

It is assumed that professor performance is measured by using a latent variable, that is, a variable which never can be measured directly. From a general statistical perspective, latent variables can be considered random effects (Bartholomew et al., 2011). On the other hand, 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 the 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 BMF 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 $\boldsymbol{\beta}_j' = (\beta_{j1}, \ldots, \beta_{jK_j-1})$ be cut points to the item j associated with each response category. Let Y_{ijs} be the rating that the is-th student assigns to professor i for the item j. Thus, the conditional probability that $[Y_{ijs} = k|\theta_i, \gamma_{is}, \boldsymbol{\beta}_j]$ is given by

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

where $k = 1, \cdots, K$, and, $\eta_{is} = \theta_i - \gamma_{is}$. The model defined in the equation (1) is not identifiable as it is common in response models. To set a scale, we assume that $\theta_i \sim N(0, 1)$. In addition, the following prior distributions are assigned.

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

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, 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 professors performance in the classroom. Therefore, it is expected that each one of the questions is answered independently by the student.

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

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

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

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

The model for evaluating the subject's relevance is the same. In this case, θ's are the relevance measures of the subjects, and the items are those designed to measure the subject relevance. Even though the students are the same, the measure of student severity changes in each case.

5 Estimation

The models were estimated using Stan (Carpenter et al., 2017), a probabilistic programming language currently used to run very complex Bayesian codes at high performance. It is easy to program as the reader can see in Appendix A. There are interfaces to run Stan using different languages like Julia, R (RStan, 2017) and others. In this study, the codes were run using RStan. Stan uses a just-in-time compiler. That is, the first time a program is run, the code is compiled. The compiler translates the Stan code to C++ and then compiles the C++ code. To implement the Lagrangian transformations required to transformed parameters, Stan uses a native automatic differentiation procedure (Bucker et al., 2006). Stan is a professional free software which implements Hamiltonian MCMC samplers (Neal, 2011) and the No-U-turn sampler (Hoffman and Gelman, 2014). To read about the implementation of general item response models in Stan, see Luo and Jiao (2017).

The original data is structured as an $M \times (p + 2)$ matrix, where M is the number of surveys, and p is the number of items in the survey. Columns 1 to p contain the scores that the student assigns to the professor for each item. Those scores are between 1 and K_j, for each column. Column $p + 1$ is the professor's *ID*, and column $p + 2$ is the student's *ID*. Since the students are anonymous, the student's *ID* is generated in a preliminary step. This *ID* is useful to work with the student's severity for the purpose of further complementary analysis.

For easy data handling and to avoid missing data, it is then re-coded as follows. The re-structured data is an $N \times 4$ matrix, where N is the total number of responses. In the case of a professor's performance measurement, column 1 is the response value, column 2 corresponds to the professor's *ID*, column 3 is the item's *ID* for the current response, and column 4 is the student's *ID*. In the case of a subject's relevance measurement, column 2 is the subject's *ID*.

Given this data arrangement, missing data is not a problem because non-responses are omitted. The Stan code for the professors performance measurement is shown in Appendix C. The code for subject relevance measurement only differs in line 15, where the latent trait is the subject relevance; in other words, it is basically the same code. The complete data and codes can be found as supplementary material at https://github.com/sicsrg/.

Appendix C shows the main components of a Stan program. Any Stan program is split into blocks. The blocks are: *data, transformed data, parameters, transformed parameters, model*, and *generated quantities*. In the current code, the *transformed data* and *transformed parameters* blocks are not required.

Block *data* contains the definition of the input data. The key words *lower* and *upper* are used to define constraints in the data, which are verified at reading time. For example, in line 03 the minimum value of N is 1. In section *parameters* the parameters to be estimated are declared. Furthermore, restrictions are permitted. Observe, for example, that in line 19, the parameter σ_β is declared positive. In this case, the restriction operates at sampling time. Stan automatically transforms the constrained variables in such a way that the internal variables have an unconstrained domain.

Stan calculates the Lagrangian factors required for the transformation of the parameters and adds them to the log-posterior. The same occurs, with the parameters defined in the *transformed parameters* section. In block *model*, the Bayesian model is defined in a declarative way. This block is similar in BUGS (Spiegelhalter et al., 1997). In addition, in the block *generated quantities*, some derived quantities can be computed. This block is useful, for example, to compute goodness of fit statistics and in general any derived functional. In the current code, the mean of the β-parameters was computed for each item in this block.

Appendix B illustrates how to run Stan code from R, using the library *rstan* (RStan, 2017). Line 04 is required to use the detected cores in the computer, to work in parallel. In line 22, the program is called to run only four iterations in order to compile the code. After that, in line 24, the compiled code is used by Stan to run the procedure.

6 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**. 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. The first task is to verify that the data from the survey are really unidimensional. In section 6.1, 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 and 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. Tools to evaluate independence in BFM models are introduced in sections 6.2 and 6.3. 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 6.5, an original GoF statistics is proposed to do that.

6.1 Analysis of unidimensionality

The first principle to define a scale to measure a unique construct is that the resulting data be unidimensional. Several authors have been working in dimension analysis of SET data. Apodacaa and Gradb (2005) used factorial techniques to review data dimensions and to perform the data analysis. Unidimensionality is the first assumption to be verified.

There are several tools to evaluate the unidimensionality of data. Three of them were used in this study: principal component analysis (PCA) (Jollife, 2002), Cronbach's alpha coefficient (Cronbach, 1951; Lord and Novick, 2013), and Cronbach Mesbah curve (CMC) (Cameletti and Caviezel, 2015).

Cronbachs alpha reliability 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 (Lord and Novick, 2013). There are different interpretations about the dimensions of the data based on the plot of the eigenvalues. In general, a first eigenvalue that is very large compared to the other values suggests unidimensionality. From this perspective, the left panel of Figure 2 suggests that the data seems unidimensional. On the other hand, the right panel of the figure shows the CMC. This curve 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 and Caviezel, 2015).

6.2 Kendall correlation

By nature, SET data are not continuous but ordinal. The associations between the variables in the model are not necessarily linear, so it is necessary to have more realistic measures than those of Person's correlation. Spearman's correlation (Spearman, 1904) is a measure that is frequently used with ordinal data and to search for non-linear relationships in the data. Nevertheless, this is a controversial measure, because it is based on the assumption that ranks are continuous variables and that Person's correlation based on ranks is valid. We recommend the Kendall rank correlation coefficient, commonly referred to as Kendall's tau (τ) coefficient (Kendall, 1938). Intuitively, the Kendall correlation between two variables will be high when observations have a similar rank and low when observations have a dissimilar rank between the two variables. From the Bayesian perspective, the

Kendal's tau is computed for each sample from the estimation process. Table 1 shows the Bayesian version of Kendall's tau for some variables in the professor's performance estimation.

6.3 Mutual information

Among the measures of independence between random variables, mutual information (MI) is singled out by its information theoretic background (Cover and Thomas, 1991; Ash, 1990). MI is zero if and only if the two random variables are strictly independent. On the other hand, estimating MI is not always easy. For this paper the *infotheo* package of R (Meyer, 2015) was used for generating MI at each Bayesian iteration. Typically, one has a set of N bivariate measurements, $z_i = (x_i, y_i), i = 1, ...N$, which are assumed to be iid (independent identically distributed) realizations of a random variable $Z = (X, Y)$ with density $f(x, y)$. Here, x and y can be either scalars or could be elements of some higher dimensional space. In the following, it is assumed that the density is a proper smooth function, although singular densities could be allowed. All that is needed is that the integrals written below exist in some sense. In particular, it is always assumed that $0 \log(0) = 0$, i.e. no need to assume that densities are strictly positive. The marginal densities of X and Y are $f_x(x) = \int f(x, y)dy$ and $f_y(y) = \int f(x, y)dx$. The MI is defined as

$$I(X, Y) = \int \int f(x, y) \log \frac{f(x, y)}{f_x(x) f_y(y)} dxdy$$

The base of the logarithm determines the units in which information is measured. In particular, using base two is equal to the amount of information measured in bits. In what follows, we will always use natural logarithms. In such case, the information is measured in nats.

6.4 Normalized Mutual Information

In order to have an adimensional measure between 0 and 1, several normalized versions of mutual information have been introduced Kavalseth (2017). The entropy or uncertainty of a random variable X is defined as

$$H(X) = -\int f_x(x) \log f_x(x) dx.$$

The normalized mutual information (NMI) that we used is given by

$$\kappa(X, Y) = \frac{I(X, Y)}{\sqrt{H(X)H(Y)}}. \quad (4)$$

.

6.5 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 goodness of fit (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 $g(\boldsymbol{W},\boldsymbol{\theta})$. Following the ideas of Box (1980), Gelfand et al. (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 idea of using discrepancy measures is studied in deep by Gelman et al. (1996). The constant 2 in the statistic is obviously redundant. It was included in order to have a counterpart to the log likelihood statistics commonly used in frequentist Statistics.

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.

6.6 Goodness of fit for Bayesian Multi-faceted models

In item response models, it is common to assess the fit of the model to the complete data and the fit of the model to items and people separately. Consequently, the adequacy of models is assessed for the complete data, the data of each one of the evaluated professors and each one of the items. 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

$$f(\boldsymbol{\omega}_{i..}|\boldsymbol{y}_{i..}) = \int f(\boldsymbol{\omega}_{i..}|\theta_i,\boldsymbol{\gamma}_i,\boldsymbol{\beta})\pi(\theta_i,\boldsymbol{\gamma}_i,\boldsymbol{\beta}|\boldsymbol{y}_{i..})d\theta_i d\boldsymbol{\gamma}_i d\boldsymbol{\beta}. \quad (5)$$

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

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

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

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

6.6.1 Computational approach to compute the GoF statistics

Let $(\boldsymbol{\theta}^{(t)},\boldsymbol{\beta}^{(t)},\boldsymbol{\gamma}^{(t)})$ be the complete t-th sample from the posterior distribution $\pi(\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\gamma}|\boldsymbol{y})$, 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 $\{\boldsymbol{\omega}_{ijs}; j = 1,\cdots,p; s = 1,\cdots,n_i\}$ from the ordered logistic distributions given by

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

   $k = 1,\cdots,K_j$. Remember that $\beta_{j0} = -\infty$, and $\beta_{jK_j} = \infty$.
2. The discrepancy measure for $\boldsymbol{\omega}_{i..}$ is given by

$$D_i(\boldsymbol{\omega}_{i..}|\theta_i^{(t)},\boldsymbol{\gamma}_i^{(t)},\boldsymbol{\beta}^{(t)}) = -2\sum_{j=1}^{p}\sum_{s=1}^{n_i}\sum_{k=1}^{K_j}\chi_k(\omega_{ijs})\log g_{ijs}(k|\theta_i^{(t)},\gamma_{is}^{(t)},\boldsymbol{\beta}_j^{(t)}).$$

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

$$D_i(\boldsymbol{y}_{i..}|\theta_i^{(t)},\boldsymbol{\gamma}_i^{(t)},\boldsymbol{\beta}^{(t)}) = -2\sum_{j=1}^{p}\sum_{s=1}^{n_i}\sum_{k=1}^{K_j}\chi_k(y_{ijs})\log g_{ijs}(k|\theta_i^{(t)},\gamma_{is}^{(t)},\boldsymbol{\beta}_j^{(t)}).$$

4. The GoF statistic is given by

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

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. In this case, The discrepancy measure for $\boldsymbol{\omega}_{.j.}$ is given by

$$D_j(\boldsymbol{\omega}_{.j.}|\boldsymbol{\theta}^{(t)},\boldsymbol{\gamma}^{(t)},\boldsymbol{\beta}_j^{(t)}) = -2\sum_{i=1}^{N}\sum_{s=1}^{n_i}\sum_{k=1}^{K_j}\chi_k(y_{ijs})\log g_{ijs}(k|\theta_i^{(t)},\gamma_{is}^{(t)},\boldsymbol{\beta}_j^{(t)}),$$

and the GoF statistics is

$$p_j = \frac{1}{T}\sum_{t=1}^{T} 1_{(D_j(\boldsymbol{\omega}_{.j.}|\boldsymbol{\theta}^{(t)},\boldsymbol{\gamma}^{(t)},\boldsymbol{\beta}_j^{(t)}) - D_j(\boldsymbol{y}_{.j.}|\boldsymbol{\theta}^{(t)},\boldsymbol{\gamma}^{(t)},\boldsymbol{\beta}_j^{(t)}))}.$$

Finally, the discrepancy measure for $\boldsymbol{\omega}$ is given by

$$D(\boldsymbol{\omega}|\boldsymbol{\theta}^{(t)},\boldsymbol{\gamma}^{(t)},\boldsymbol{\beta}^{(t)}) = -2\sum_{i=1}^{N}\sum_{s=1}^{n_i}\sum_{j=1}^{p}\sum_{k=1}^{K_j}\chi_k(y_{ijs})\log g_{ijs}(k|\theta_i^{(t)},\gamma_{is}^{(t)},\boldsymbol{\beta}_j^{(t)}),$$

7 Real Case Data

The data come from a student survey that runs almost every semester in the Natural Sciences Faculty at the Universidad Nacional de Colombia. The survey is designed to measure the professors' performance in the classroom, the subject relevance of the course, and the efficiency of the available resources on the campus. The measurement is run from the students' perception. In this study, the data from the second semester of 2015 were used. The original data had 14703 responses (surveys), which includes 606 courses, 306 subjects, and 401 professors. For ease, we selected a sample of 1380 surveys, which included 124 professors and 123 subjects. The students came from 36 programs, including both undergraduate and postgraduate students. The selected subjects in the sample had between 4 and 20 surveys, and the professors have between 3 and 36 surveys each. The difference in the lower limit was due to shared courses in which not all professors were scored. Appendix F and Figure 3 show the main results for the professors' performance measurement. On the other hand, Appendix E and Figure 4 show the results of each subject's relevance measurement. The questionnaire, complete data, sample data, and all the codes can be found as supplementary material.

7.1 Unidimensionality Analysis

The left panel of Figure 2 shows the box plot of the eigenvalues of the PCA analysis of the professors' performance data, assuming (carefully) that the values are real. In reality, they are ordinal. According to the PCA, the first component explains 41.2% of the variance in the data. From this perspective, the data seems unidimensional. On the other hand, the right panel of the figure shows the CMC. The unique item that seems problematic is item 23 (9 in Appendix A). This item could be taken out, since it might not belong to the *professors' performance in the classroom*. However, in the estimation analysis it was included because it is not really problematic, and it is useful for further discussion. In addition, Cronbach's alpha reliability for the professor's performance data was 0.885. Thus, it is possible to run the estimation of the model to obtain the professors' performance measurement. The analysis of unidimensionality for the subject's relevance was done in a similar way.

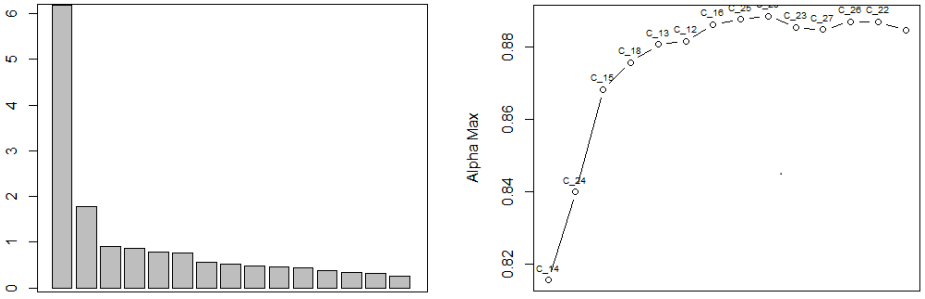

Fig. 2 Barplot of the eigenvalues of PCA analysis and Cronbach-Mesbah Curve

7.2 Professor's performance estimation

The data to measure professor's performance included 15 items. Appendix A shows the questions designed to measure each professor's performance. Each item had 5 categories originally; however, after a preliminary analysis, it was found that the lowest category had very few responses. Table 4 in section 7.4 shows that the self-performance is concentrated in the values 3, 4, and 5. This fact may be related to the low frequency of responses in the lower categories. Consequently, the first two categories were merged, and the items were re-coded to have four categories.

Once the data was restructured, as explained in section 5, it became a 20617×4 matrix, including 15 items, 1380 surveys, and 124 professors. To run the Stan code, the number of iterations was set to 2000, and 4 chains were requested per run for both professor's performance and subject's relevance. The first 1000 iterations were discarded as a burnout period. In total, 4000 samples from the posterior distribution were obtained for each case. See line 24 in Appendix D.

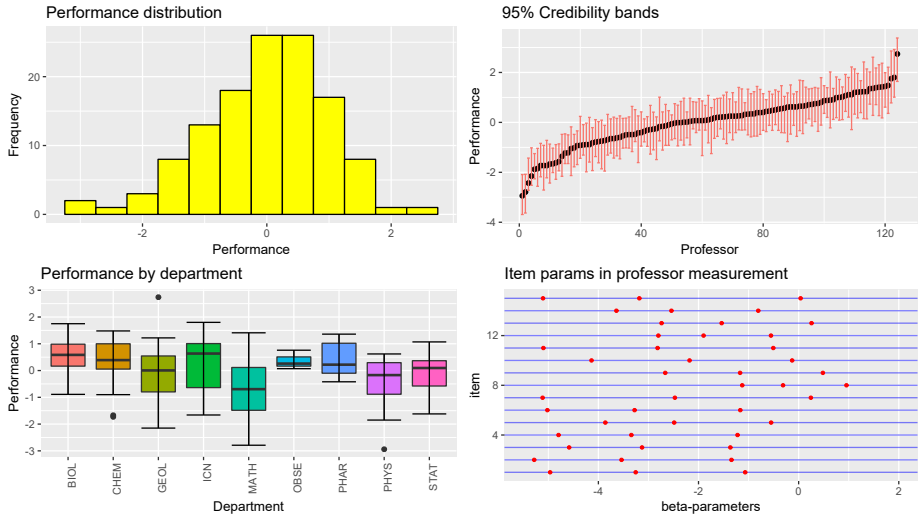

Fig. 3 Plots from professor's performance estimation computed from the Stan output.

The table in Appendix F shows the estimates computed by Stan. The $\hat{R}$ statistic for the convergence diagnostic column (Gelman and Rubin, 1992) was omitted. All their values were very close to 1, so all the chains were theoretically convergent. Figure 3 shows the main results of the professor performance estimation.

The top left panel is the plot of the estimate latent trait: the professor's performance. A priori, it was supposed that the latent trait was a random variable distributed as an $N(0, 1)$. The top right panel shows the estimation of the performance of all professors, including the corresponding 95% credibility bands. The professors were ordered according to their measure. The bottom left panel is the distribution of the latent trait by departments. The comparisons by departments

are valid since the measurements are actually real numbers. Finally, the bottom right panel show a general image of the test. Each line represents an item, and the red dots are the cut-points of the items.

It can be interpreted that item 2 was the easiest item, in the sense that, in general, students tended to rate the professors highly. The question in this case was: Did the professor respect the agreed dates for academic activities, including evaluations and delivery of results? In contrast, item 9 was the most difficult, in the sense that the students gave the professors lower scores. In this case, the question was: Does the professor organize activities that allow you to exercise oral and written expression? This item was reported in section 6.1 as an item that possibly does not belong to the *professor's performance measure in the class room.*

Appendix F shows that the generated quantities μ_β are useful to perform these analyses. In the current case, $\mu_{\beta_2} = -3.39$ and $\mu_{\beta_8} = -0.16$ are the two extreme values. Remember that the μ_{β_k}-values are estimates of the mean of the $\mu_{\beta_{kj}}$ cut-points of the items. Thus, on average, item 2 can be considered easier and item 8 more difficult. In addition, it is important to note that the standard deviation of the β_{jk}'s distributions are really small due to the large sample for the items. Consequently, the estimates of such parameters include only a small uncertainty.

params/stats	q2.5	q25	q50	mean	q75	q97.5
(θ, γ)	-0.047	-0.022	-0.010	-0.010	0.003	0.026
(θ, ρ)	0.735	0.768	0.785	0.784	0.800	0.831
(θ, ξ)	0.140	0.158	0.167	0.167	0.176	0.192
(γ, ρ)	-0.112	-0.087	-0.073	-0.074	-0.061	-0.037
(γ, ξ)	-0.202	-0.185	-0.176	-0.176	-0.167	-0.149

Table 1 Bayesian estimation of Kendall's correlation estimation in the professor's performance estimation. The variables are: θ: professor's performance, γ: student's severity, ρ: harmonic mean, ξ: student's self-performance.

params/stats	q2.5	q25	q50	mean	q75	q97.5
(θ, θ)	1.600	1.606	1.607	1.607	1.609	1.610
(θ, γ)	-0.004	-0.002	0.0021	0.002	0.0043	0.009
(θ, ρ)	0.602	0.668	0.707	0.709	0.747	0.834
(θ, ξ)	0.019	0.023	0.026	0.026	0.028	0.033
(γ, γ)	2.401	2.401	2.401	2.401	2.401	2.401
(γ, ρ)	-0.002	0.002	0.005	0.005	0.007	0.015
(γ, ξ)	0.019	0.025	0.028	0.028	0.031	0.037
(ρ, ρ)	.610	1.610	1.610	1.610	1.610	1.610
(ρ, ξ)	0.029	0.029	0.029	0.029	0.029	0.029
(ξ, ξ)	0.969	0.969	0.969	0.969	0.969	0.969

Table 2 Bayesian estimation of mutual information in the professor's performance estimation. The variables are: θ: professor's performance, γ: student's severity, ρ: harmonic mean, ξ: student's self-performance.

Tables 1, 2, and 3 show the Bayesian estimation of Kendall's correlation, mutual information, and normalized mutual information for the variables of interest. The

params/stats	q2.5	q25	q50	mean	q75	q97.5
(θ, γ)	-0.002	-0.000	0.001	0.001	0.002	0.005
(θ, ρ)	0.369	0.410	0.432	0.433	0.456	0.503
(θ, ξ)	0.015	0.019	0.021	0.021	0.022	0.026
(γ, ρ)	0.001	0.001	0.002	0.003	0.004	0.007
(γ, ξ)	0.013	0.016	0.018	0.018	0.020	0.024
(ρ, ξ)	0.023	0.023	0.023	0.023	0.023	0.023

Table 3 Bayesian estimation of normalized mutual information in the professor's performance estimation. The variables are: θ: professor's performance, γ: student's severity, ρ: harmonic mean, ξ: student's self-performance.

harmonic mean of the professors' scores were included only for comparison. As expected, professor's performance and student's severity were independent variables. On the other hand, there was a small positive association between student-reported self-performance and professor's performance and a small negative association between reported student's self-performance and student's severity.

7.3 Subject relevance estimation

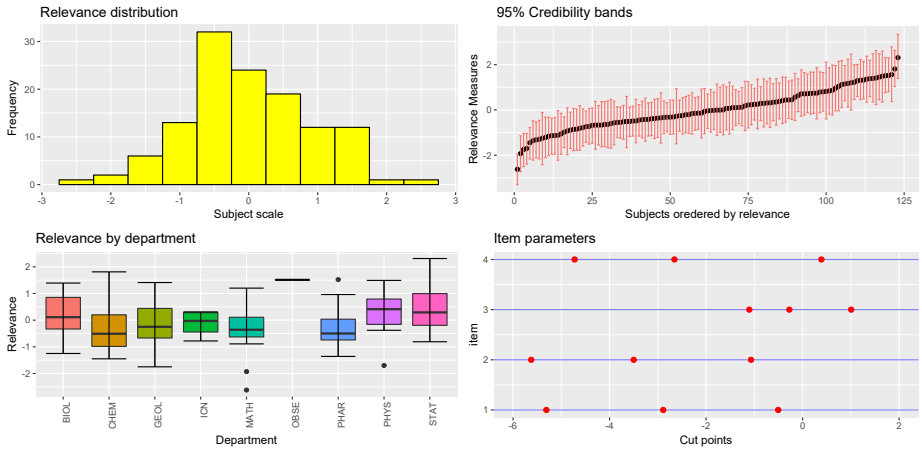

Fig. 4 Plots from subject relevance estimation computed from the Stan output.

In this case, the data is a 5492×4 matrix, including four items with four categories after merging categories 1 and 2, 1380 surveys, and 123 subjects. Figure 4 shows the main results of the subject relevance estimation; Appendix E includes the estimate parameters, and Appendix B contains the questions designed to measure subject relevance. It should be noted that in this case $\sigma_\beta = 2.62$ is almost twice the corresponding value in the professor performance estimate. This is the effect of the number of items used in this case. Nevertheless, these items are more precise because they have more categories. On the other hand, the severitys standard deviation is $\sigma_\gamma = 1.79$, which is also greater than 1.20 reported in the professor's performance measurement.

From a statistical standpoint, the interpretations are similar to those in the previous section. However, in this case, the latent trait is very different. The results reflect the perception that the students have about the subjects they are studying. In terms of evaluation of academic programs, this information is really useful. In the current case, the information about item 3 confirms that the majority of the students consider that some of the subjects are not useful for their professional future. The question in this case was: Do you consider that the theoretical and practical contents of the subject are useful for professional activity? This result may be associated with several factors, such as the students' experience, their particular interests, and so on.

7.4 Severity analysis

The inclusion of the severity/leniency parameters is very important in order to correct some of the problems associated with using surveys in measuring professor performance. Figure 5 shows the main information that comes from the severity estimation. The top left panel in the figure shows the distribution plot of the estimate of severity among the 1380 students computed from the professor's performance estimation. The plot on the top right shows that the severity changed between the professor's performance measurement and the subject's relevance measurement. As mentioned before, the standard deviations are different. Additionally, the correlation between the latent traits estimated for both cases was 0.65.

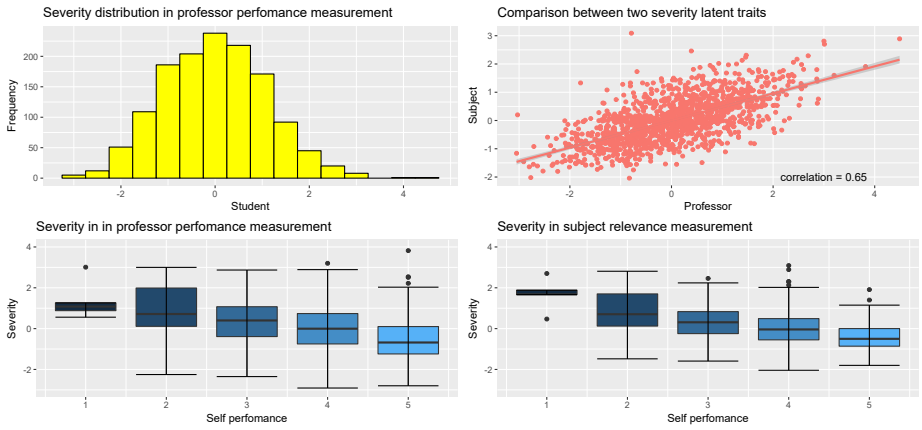

Fig. 5 Severity Analysis.

In the survey, all the students were anonymous. However, some particular socio-demographic questions about the students were included: sex, field of study, number of semesters completed, number of times the student failed the subject, and a self-evaluation of their performance in the course. This information was used to find relationships with the estimate severity latent trait. The only relationship found was with the self-evaluation. The scale in this case was from 1 to 5, and

this self-evaluation was found to be a good approximation of the score expected by the student in the course. Table 4 shows the distribution of the self-performance variable. The values 1 and 2 had a low frequency, as mentioned above.

Value	Frequency
1	5
2	22
3	246
4	890
5	217

Table 4 Distribution of the self-performance variable in the sample

Similarly, as expected, an inverse relationship was found between the self-evaluation and the severity of the student: the more severe the student, the worse the self-evaluation. However, the association between variables in these cases was really small in the data.

7.5 Strategic diagram

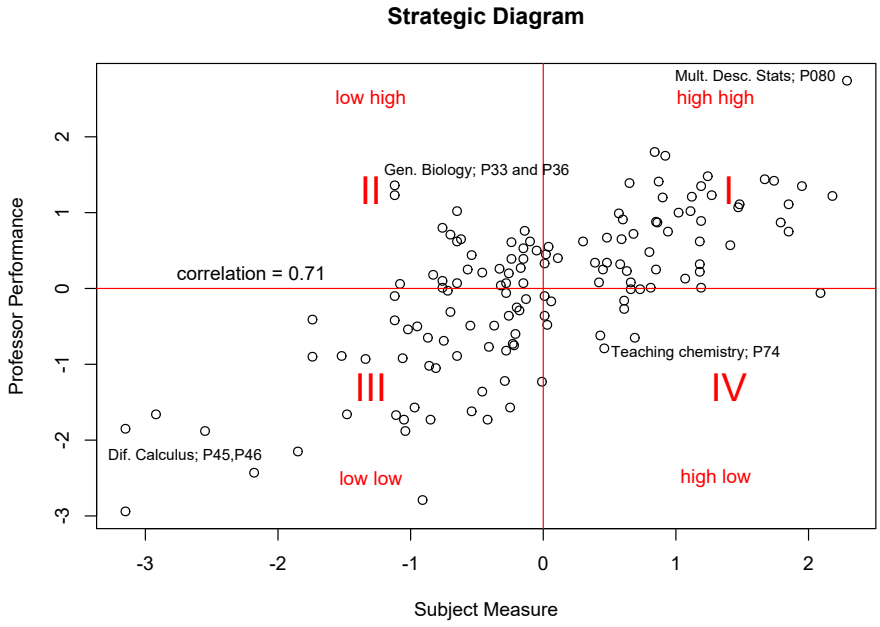

Fig. 6 Strategic diagram

Figure 6 shows the strategic diagram used in different areas as a decision-making aid. In the diagram, the latent traits of professor performance vs. subject relevance

were plotted. It is quite interesting to observe how the professors and subjects are organized. For example, the professors with the highest performance and the most relevant subjects are located in the top right quadrant. This diagram allows for a variety of interesting interpretations. Furthermore, the diagram is a tool to select professors and subjects for further, more detailed studies. For example, one might wonder: What do the best professors in the more relevant subjects do differently than lower rated professors?

7.6 Goodness of fit

The file *R40 GoF* in the supplementary material shows the implementation of the Gof tests. The value of the global GoF statistics was 0.43., so the model fits the complete data nicely. On the other hand, some items and professors have values that are very low or very high. Complete information can be consulted in files *pval g.txt*, *pval item.txt*, and *pval professor.txt*. This information is useful to review the results for the purpose of improving the instruments.

8 Discussion and Recommendations

SET information is commonly used by a lot of universities as an aid in decision-making processes. However, the majority of reported measures based on SET data have problems in consistency and mathematical validity. The main problem is that the data coming from SET surveys is categorical, but most of the time is treated as if it were numerical. From this perspective, the research based on this type of measure is not reliable.

In general, the mathematical manipulation could be valid only in the case of dichotomous variables, as long as the variables represent the presence or absence of a characteristic. Nevertheless, in this case, the classical approach does not permit assessing the reliability of the measures.

In this paper, a sophisticated statistical instrument was introduced to measure the performance of the professor in class and the relevance of the subject based on the information SET coming from student surveys.

A complete statistical model was proposed to set up accurate and real measures of the performance of the professors. Obviously, these measures are not perfect. However, they appear to be mathematically valid and have a strong theoretical support.

A family of models was proposed, since the linear latent predictor may be modified to consider those variables that experts consider more influential in the performance measure. For example, some studies have reported that a professor's attractiveness might influence the measurement. In this case, a cross variable of the students gender and professors gender, along with the gender variable, may be included in the predictor to isolate the effect. In the real case data example, it was included gender variables and other variables. However, none of them were

statistically different from zero. For all of those variables, the credibility interval included zero.

The distribution of the self-performance variable and student severity also deserves special attention. Some researchers have found high positive correlations between this variable and the scores that the students assign to the professors. In the data used in this research, it was found that self-performance was concentrated mainly in the value 4 from the scale defined by the ordinal categories 1, 2, 3, 4, and 5. As can be seen in the bottom left panel of Figure 5, for category 4, the severity is centered at zero and distributed as a normal variable. The distribution of severity in categories 3 and 5 is a little displaced from zero, but taking similar values. These findings confirm that, in general, there was a tendency among students to be more severe when the student expected a low score and vice versa. Nevertheless, the severity variable introduced in this research is distributed along the real numbers, and severity is only marginally related to expected scores. According to the data, students with a low expected score tend to be more severe, although with differing levels of severity. On the other hand, students with high expected scores tend to be less severe, although the effect in this case is weaker.

As previously mentioned, the inclusion of severity parameters in statistical models for using in evaluation processes is not new. However, the models proposed in this study are different. Classical multi-faceted models are Rasch models, while in this work, multi-faceted Bayesian models are proposed. In classical models, the raters are few and are well-trained. In the proposed models, the raters are students, and in general, there are many of them.

In using the BMF models, the objective was to measure a latent trait of the professors and their performance in the classroom based on student responses, under the assumption that the students may have a biased opinion about the performance of their professors. This bias was modeled by the severity parameter. The main assumption was that a fair response to the survey implies that the student has an unbiased opinion, so his/her severity parameter must be zero. It was assumed that the severity is a random variable with normal distribution centered at zero.

Figure 5 confirms the findings in the literature in the sense that the students' responses were influenced by their expected scores in the course. The inclusion of the student severity parameter in the statistical models tries to alleviate this problem in the performance measurement. The key feature of the proposed models is that they are designed to filter the information received from students about the performance of professors in order to correct potential biases. In this sense, several issues must be taken into account.

First, point and interval estimates are calculated. This is a very significant characteristic because interval estimates are more realistic in this type of measurement. Figure 3 illustrates the fact that close point estimates are indistinguishable and that measurement error is not too small. Thus, it is not advisable to make decisions based solely on the ordered performance of the professors. The usefulness of these measures is the ability to distinguish between professors with low, medium and high performance. Second, it is very important to validate the measurement instruments. Frequent problems arise when the questions are not well formulated. The GoF

statistics introduced in section 6.5 are very useful for evaluating the instrument. An item with a statistical value GoF very close to 0.0 or 1.0 is problematic and should be reviewed. In this sense, it is convenient to consider standardized questions proposed in the literature. In addition, it is very important to define appropriate scales with the same number of levels of response categories and try to train students in the manner they must respond. Professors should encourage the most equitable responses possible for the survey. Third, problematic GoF statistics for professor measurements may be caused by either very divergent or few responses. These cases may be isolated to verify the problem in order to take it into account within the final reports.. Fourth, the additional information available in the survey from students and professors must be included in preliminary models. This is an important preliminary step. If a variable is found to be significant in the model for a particular application, its inclusion in the final model improves the performance estimates.

Finally, the strategic diagram introduced in section 7.5 is a valuable tool in the academic evaluation of the courses and as a decision-making aid.

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 in collaboration with the Statistics Department at the Universidad de Valparaíso.

References

Aigner, D. and Thum, F. (1986). On student evaluations of teaching ability. *Journal of Economic Education*, 17 (Fall):243--266.

Apodacaa, P. and Gradb, H. (2005). The dimensionality of student ratings of teaching: integration of uni- and multidimensional models. *Studies in Higher Education*, 30(6):723--748.

Ash, R. (1990). *Information Theory*. Dover.

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

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

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

Becker, W. and Watts, M. (1999). How departments of economics should evaluate teaching. *American Economic Review (Papers and Proceedings)*, 89:344349.

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

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

Box, G. (1980). Sampling and bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society. Series A.*, 143:383--430.

Braga, M., Paccagnella, M., and Pellizzari, M. (2014). Evaluating students evaluations of professors. *Economics of Education Review*, 41:71--88.

Bucker, M., Corliss, G., Hovland, P., Naumann, U., and Norris, B., editors (2006). *Automatic Differentiation: Applications, Theory, and Implementations*. Springer.

Cameletti, M. and Caviezel, V. (2015). *Package CMC. CRAN R-project*.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, Articles*, 76(1):1--32.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. Wiley.

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

Decanio, S. (1986). Student evaluations of teaching a multinominal logit approach. *Journal of Economic Education*, 17 (Summer):165176.

Ditts, D. (1980). A statistical interpretation of student evaluation feedback. *Journal of Economic Education*, 11(2):10--15.

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

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

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

Gelfand, A. E., Dey, D., and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. Technical report, DTIC Document.

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6(4):733--760.

Gelman, A. and Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statist. Sci.*, 7:457--472.

Guler, N. (2014). Analysis of open-ended statistics questions with many facet rasch model. *Eurasian Journal of Educational Research*, 2014:73--90.

Hoffman, M. and Gelman, A. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15.

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

Kavalseth, T. O. (2017). On normalized mutual information: Measure derivations and properties. *Entropy*, 19:631--644.

Kendall, M. (1938). A new measure of rank correlation. *Biometrika*, 30:81--89.

Krautmann, A. C. and Sander, W. (1999). Grades and student evaluations of teacher. *Economics of Education Review*, 18:59--63.

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

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

Luo, Y. and Jiao, H. (2017). Using the stan program for bayesian item response theory. *Educational and Psychological Measurement*, 2017:1--25.

Mehdizadeh, M. (1990). Loglinear models and student course evaluations. *Journal of Economic Education*, 21 (Winter):7--21.

Meyer, P. E. (2015). *Package Infotheo. CRAN R-project*.

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

Peters, A., Sahloff, E., and Stone, G. (2010). Instructional design and assessment. a standardized rubric to evaluate student presentations. *American Journal of*

Pharmaceutical Education, 74(9), Article 71:1--8.

Pin, S., Chen, P. H., and Chen, H. (2012). Improving creativity performance assessment: A rater effect examination with many facet rasch model. *Creativity Research Journal*, 24:345--357.

RStan, D. T. (2017). *RStan: the R interface to Stan. R package version 2.16.2.*

Seiver, D. (1983). Evaluations and grades: a simultaneous framework. *Journal of Economic Education*, 14 (Summer):32--38.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15:72--101.

Spiegelhalter, D. J., Thomas, A., Best, N. G., and Gilks, W. (1997). *BUGS: Bayesian inference using Gibbs sampling, Version 0.60.* Cambridge: Medical Research Council Biostatistic Unit.

Stark, P. and Freishtat, R. (2014). An evaluation of course evaluation. *ScienceOpen Research- Section SOR-EDU*, pages 1--7.

Uttl, B., White, C., and Wong, D. (2017). Meta-analysis of facultys teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, 54:22--42.

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

Wolfe, E. and Chiu, C. (1997). Detecting rater effects with a multi-faceted rating scale model.

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 teacher 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 teacher's performance was:

B Questionnaire used to Measure Subject Relevance

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

1. Do you consider the content of the subject to be related to that of other subjects?
2. Do you consider the theoretical and practical content of the subject to be useful for the professional activity?
3. Did the topics seen in the course meet your expectations?
4. How would you rate the subject in general terms?

C Stan Code for Estimating Professor Performance

```
01
02  data{
03  int<lower=1>  N;                        // number of responses
04  int<lower=10> N_prof;                   // number of professors
05  int<lower = 10> N_stud;                 // number of students
06  int<lower=2>  N_item;                   // number of items
07  int<lower=2,upper=5> N_cat;             // number of categories
08  int<lower=1,upper=N_cat> y[N];          //  y[n], n-th response
09  int<lower=1,upper=N_prof> professor[N]; // professor of response[n]
10  int<lower=1,upper=N_item> item[N];      // item of response[n]
11  int<lower=1,upper=N_stud> student[N];   // professor of response[n]
12  }
13
14 parameters{
15  vector[N_prof] theta;          // latent traits of professors
16  vector[N_stud] gamma;          // students'severity
17  ordered[N_cat-1] beta[N_item]; // beta parameters
18  real mu_beta;                  // mean of the beta-parameters
19  real<lower=0> sigma_beta;      // sd of beta prior distributions
20  real<lower=0> sigma_gamma;     // sd of gamma prior distributions
21  }
22
23 model{
24  theta ~ normal(0,1);             // scale of the latent traits
25  gamma ~ normal(0,sigma           //scale of the students'severity
26  for(j in 1:N_item){
27   beta[j]~ normal(mu_beta,sigma_beta); // prior of betas
28   }
29  mu_beta ~ normal(0,2);     // hyper prior for mu_beta
30  sigma_beta ~ cauchy(0,2);  // hyper prior for sigma_beta
31  sigma_gamma ~ cauchy(0,2); // hyper prior for sigma_gamma
32  //likelihood
33  for(n in 1:N){
34   y[n]~ ordered_logistic(theta[professor[n]]-gamma[student[n]],beta[item[n]]);
35   }
36 }
37
38 generated quantities {
39  vector[N_item] mu_item_beta; // the mean of betas by item
40  for (n in 1: N_item){
41    mu_item_beta[n] = mean(beta[n]);
42   }
```

D RStan Code for Estimating Professor Performance

```
01
02 library(rstan)
03 rstan_options(auto_write = TRUE)
04 options(mc.cores = parallel::detectCores())
05
06 # Read the data
07 setwd("D:/Alvaro/Alvaro_2018/Proyecto_Edificando")
08 load(file="professor_data_sample.Rdata")
09
10 # total parameters
11 NN = nrow(professor_data) #20617; num. registers
12 NS = length(unique(professor_data[,2])) #1380; num. students
13 NP = length(unique(professor_data[,3])) #124;  num. professors
14 NI = length(unique(professor_data[,4])) #15;   num. items
15 NC = max(professor_data[,1]) #3; number of categories
16 #
17 # go to Stan
18 dat = list(professor=professor_data[,3], student = professor_data[,2],
   item = professor_data[,4], y = professor_data[,1], N=NN, N_item=NI, N_prof=NP,
   N_cat=NC, N_stud=NS)
19
20 #first compile and save the fitted model for re-using
21 # data are taken from the current R works
22 prof_fit_p_ <- stan(file = 'Multi_Faceted_Ciencias_AM_002.stan', data = dat,
   iter = 4, chains = 1)
23 # now sample using the compiled model
24 prof_fit_2<- stan(fit = prof_fit_p_,  data =dat, iter = 2000, chains = 4,
   control = list(max_treedepth = 12))
25 # save the stan object
26 save(prof_fit_2,file="Model_2_prof_fit_2_pr_it_st.Rdata")
```

E Estimated Item Parameters in Subject Relevance Measurement

Param	mean	sd	q2.5	q25	q50	q75	q97.5	n eff
$\beta_{1,1}$	-6.26	0.27	-6.80	-6.44	-6.25	-6.07	-5.74	1588.99
$\beta_{1,2}$	-3.52	0.16	-3.83	-3.62	-3.52	-3.41	-3.22	1367.95
$\beta_{1,3}$	-0.68	0.13	-0.92	-0.76	-0.67	-0.59	-0.43	1604.77
$\beta_{2,1}$	-6.63	0.31	-7.25	-6.83	-6.62	-6.42	-6.04	1836.10
$\beta_{2,2}$	-4.23	0.17	-4.58	-4.35	-4.23	-4.12	-3.89	1512.71
$\beta_{2,3}$	-1.36	0.13	-1.61	-1.45	-1.36	-1.27	-1.09	1306.21
$\beta_{3,1}$	-4.70	0.18	-5.05	-4.82	-4.70	-4.57	-4.35	1197.44
$\beta_{3,2}$	-2.37	0.14	-2.64	-2.46	-2.37	-2.27	-2.10	1240.35
$\beta_{3,3}$	0.18	0.13	-0.06	0.09	0.18	0.27	0.43	1676.10
$\beta_{4,1}$	-5.67	0.23	-6.12	-5.83	-5.68	-5.52	-5.23	1781.68
$\beta_{4,2}$	-3.27	0.15	-3.56	-3.37	-3.27	-3.17	-2.98	1512.92
$\beta_{4,3}$	0.45	0.13	0.21	0.37	0.45	0.54	0.71	1789.05
μ_{β_1}	-3.48	0.15	-3.79	-3.59	-3.48	-3.38	-3.19	1157.65
μ_{β_2}	-4.07	0.17	-4.40	-4.18	-4.08	-3.96	-3.74	1295.90
μ_{β_3}	-2.30	0.13	-2.54	-2.38	-2.30	-2.21	-2.05	1078.99
μ_{β_4}	-2.83	0.14	-3.10	-2.93	-2.83	-2.73	-2.56	1367.00
σ_β	2.62	0.62	1.74	2.19	2.51	2.94	4.15	4000.00
σ_γ	1.79	0.07	1.66	1.74	1.78	1.83	1.92	787.11

Table 5 Estimate β-parameters in subject relevance estimation computed by Stan. Column Rhat ($\widehat{R}$) is omitted. All values are very close to 1.00. Column *n eff* is the effective sample size.

F Estimated Item Parameters in Professor Performance Measurement

Param	mean	se mean	sd	q2.5	q25	q50	q75	q97.5	n eff
$\beta_{1,1}$	-4.96	0.00	0.22	-5.41	-5.11	-4.96	-4.81	-4.54	1754.54
$\beta_{1,2}$	-3.25	0.00	0.15	-3.55	-3.35	-3.25	-3.15	-2.97	963.21
$\beta_{1,3}$	-1.07	0.00	0.12	-1.30	-1.15	-1.07	-0.99	-0.84	680.11
$\beta_{2,1}$	-5.28	0.01	0.23	-5.75	-5.44	-5.28	-5.12	-4.83	1751.05
$\beta_{2,2}$	-3.54	0.00	0.15	-3.83	-3.64	-3.54	-3.44	-3.24	1121.02
$\beta_{2,3}$	-1.34	0.00	0.12	-1.57	-1.42	-1.34	-1.26	-1.10	712.73
$\beta_{3,1}$	-4.58	0.01	0.19	-4.96	-4.70	-4.58	-4.46	-4.23	1120.75
$\beta_{3,2}$	-3.13	0.00	0.14	-3.40	-3.22	-3.13	-3.04	-2.86	909.86
$\beta_{3,3}$	-1.36	0.00	0.12	-1.59	-1.44	-1.36	-1.28	-1.13	689.49
$\beta_{4,1}$	-4.80	0.01	0.20	-5.20	-4.93	-4.79	-4.66	-4.41	1366.11
$\beta_{4,2}$	-3.34	0.00	0.14	-3.63	-3.44	-3.34	-3.25	-3.07	959.73
$\beta_{4,3}$	-1.22	0.00	0.12	-1.46	-1.30	-1.22	-1.15	-0.99	688.38
$\beta_{5,1}$	-3.86	0.00	0.16	-4.18	-3.97	-3.86	-3.75	-3.54	992.43
$\beta_{5,2}$	-2.49	0.00	0.13	-2.74	-2.57	-2.49	-2.40	-2.23	773.51
$\beta_{5,3}$	-0.55	0.00	0.12	-0.78	-0.63	-0.55	-0.47	-0.32	667.26
$\beta_{6,1}$	-5.02	0.00	0.20	-5.43	-5.16	-5.01	-4.88	-4.63	1402.25
$\beta_{6,2}$	-3.28	0.00	0.14	-3.55	-3.37	-3.27	-3.18	-3.01	932.93
$\beta_{6,3}$	-1.17	0.00	0.12	-1.40	-1.24	-1.16	-1.09	-0.93	692.72
$\beta_{7,1}$	-5.11	0.01	0.22	-5.57	-5.26	-5.11	-4.96	-4.69	1580.30
$\beta_{7,2}$	-2.47	0.00	0.13	-2.72	-2.56	-2.47	-2.38	-2.21	791.58
$\beta_{7,3}$	0.24	0.00	0.12	0.01	0.16	0.25	0.32	0.47	703.15
$\beta_{8,1}$	-1.13	0.00	0.12	-1.36	-1.21	-1.12	-1.05	-0.90	657.09
$\beta_{8,2}$	-0.31	0.00	0.12	-0.54	-0.39	-0.31	-0.23	-0.09	673.70
$\beta_{8,3}$	0.95	0.00	0.12	0.72	0.87	0.96	1.04	1.19	696.55
$\beta_{9,1}$	-2.66	0.00	0.13	-2.92	-2.75	-2.66	-2.57	-2.42	760.44
$\beta_{9,2}$	-1.17	0.00	0.12	-1.40	-1.25	-1.17	-1.09	-0.94	681.27
$\beta_{9,3}$	0.48	0.00	0.12	0.24	0.40	0.48	0.56	0.71	660.56
$\beta_{10,1}$	-4.14	0.01	0.17	-4.47	-4.25	-4.13	-4.03	-3.81	931.12
$\beta_{10,2}$	-2.18	0.00	0.13	-2.42	-2.26	-2.18	-2.09	-1.94	739.15
$\beta_{10,3}$	-0.13	0.00	0.12	-0.36	-0.21	-0.13	-0.05	0.10	641.42
$\beta_{11,1}$	-5.10	0.01	0.23	-5.55	-5.25	-5.09	-4.94	-4.66	1642.22
$\beta_{11,2}$	-2.82	0.00	0.14	-3.09	-2.91	-2.82	-2.73	-2.56	897.91
$\beta_{11,3}$	-0.51	0.00	0.12	-0.74	-0.59	-0.51	-0.43	-0.28	702.58
$\beta_{12,1}$	-2.80	0.00	0.14	-3.06	-2.89	-2.80	-2.71	-2.54	836.47
$\beta_{12,2}$	-1.90	0.00	0.12	-2.14	-1.98	-1.90	-1.81	-1.66	740.87
$\beta_{12,3}$	-0.55	0.00	0.12	-0.78	-0.63	-0.55	-0.47	-0.32	670.18
$\beta_{13,1}$	-2.74	0.00	0.13	-3.00	-2.83	-2.74	-2.65	-2.49	772.05
$\beta_{13,2}$	-1.54	0.00	0.12	-1.77	-1.62	-1.54	-1.46	-1.30	669.18
$\beta_{13,3}$	0.26	0.00	0.12	0.03	0.18	0.26	0.34	0.49	634.23
$\beta_{14,1}$	-3.64	0.00	0.15	-3.94	-3.74	-3.64	-3.54	-3.33	919.32
$\beta_{14,2}$	-2.54	0.00	0.13	-2.79	-2.63	-2.54	-2.45	-2.29	727.30
$\beta_{14,3}$	-0.81	0.00	0.12	-1.04	-0.89	-0.81	-0.73	-0.58	658.45
$\beta_{15,1}$	-5.11	0.00	0.21	-5.51	-5.25	-5.11	-4.96	-4.72	1722.68
$\beta_{15,2}$	-3.18	0.00	0.14	-3.47	-3.28	-3.18	-3.09	-2.91	946.17
$\beta_{15,3}$	0.04	0.00	0.12	-0.19	-0.04	0.04	0.12	0.27	642.62
μ_{β_1}	-3.10	0.00	0.14	-3.37	-3.19	-3.10	-3.00	-2.83	812.57
μ_{β_2}	-3.39	0.00	0.14	-3.66	-3.48	-3.39	-3.29	-3.12	892.02
μ_{β_3}	-3.02	0.00	0.13	-3.28	-3.11	-3.02	-2.94	-2.77	727.90
μ_{β_4}	-3.12	0.00	0.13	-3.38	-3.21	-3.12	-3.03	-2.86	787.93
μ_{β_5}	-2.30	0.00	0.12	-2.54	-2.38	-2.30	-2.22	-2.06	661.07
μ_{β_6}	-3.15	0.00	0.13	-3.41	-3.24	-3.15	-3.07	-2.90	769.64
μ_{β_7}	-2.45	0.00	0.13	-2.70	-2.53	-2.44	-2.36	-2.19	763.13
μ_{β_8}	-0.16	0.00	0.11	-0.38	-0.24	-0.16	-0.08	0.06	609.70
μ_{β_9}	-1.12	0.00	0.11	-1.34	-1.19	-1.12	-1.04	-0.90	600.30
$\mu_{\beta_{10}}$	-2.15	0.00	0.12	-2.38	-2.23	-2.15	-2.07	-1.92	604.93
$\mu_{\beta_{11}}$	-2.81	0.00	0.13	-3.07	-2.90	-2.81	-2.72	-2.56	790.75
$\mu_{\beta_{12}}$	-1.75	0.00	0.12	-1.98	-1.83	-1.75	-1.67	-1.53	655.64
$\mu_{\beta_{13}}$	-1.34	0.00	0.11	-1.56	-1.42	-1.34	-1.26	-1.12	590.41
$\mu_{\beta_{14}}$	-2.33	0.00	0.12	-2.57	-2.41	-2.33	-2.25	-2.08	647.12
$\mu_{\beta_{15}}$	-2.75	0.00	0.13	-3.00	-2.84	-2.75	-2.66	-2.50	794.83
σ_β	1.82	0.00	0.2	1.46	1.67	1.80	1.94	2.26	4000.00
σ_γ	1.20	0.00	0.03	1.13	1.17	1.20	1.22	1.27	2074.98

Table 6 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.