Nonparametric Item Response Models: A Comparison on Recovering True Scores

Vithor Rosa Franco[1]
Marie Wiberg[2]
Rafael Valdece Sousa Bastos[1]

[1]*São Francisco University, Campinas, São Paulo, Brasil*
[2]*Umeå School of Business, Economics and Statistics, Umeå University, Umeå, Sweden*

Abstract
Nonparametric procedures are used to add flexibility to models. Three nonparametric item response models have been proposed, but not directly compared: the Kernel smoothing (KS-IRT); the Davidian-Curve (DC-IRT); and the Bayesian semiparametric Rasch model (SP-Rasch). The main aim of the present study is to compare the performance of these procedures in recovering simulated true scores, using sum scores as benchmarks. The secondary aim is to compare their performances in terms of practical equivalence with real data. Overall, the results show that, apart from the DC-IRT, which is the model that performs the worse, all the other models give results quite similar to those when sum scores are used. These results are followed by a discussion with practical implications and recommendations for future studies.
Keywords: Nonparametric item response model; Bayesian modeling; Monte Carlo simulation.

Modelos de Resposta ao Item Não-Paramétricos: Comparando a Recuperação de Escores Verdadeiros

Resumo
Procedimentos não paramétricos são usados para adicionar flexibilidade aos modelos. Três modelos não paramétricos de resposta ao item foram propostos, mas não comparados diretamente: o Kernel *smoothing* (KS-IRT); a Curva Davidiana (DC-IRT); e o modelo semiparamétrico Rasch Bayesiano (SP-Rasch). O objetivo principal do presente estudo é comparar o desempenho desses procedimentos na recuperação de escores verdadeiros simulados, utilizando escores de soma como benchmarks. O objetivo secundário é comparar seus desempenhos em termos de equivalência prática com dados reais. De forma geral, os resultados mostram que, além do DC-IRT, que é o modelo que apresenta o pior desempenho, todos os outros modelos apresentam resultados bastante semelhantes aos de quando se usam somatórios. Esses resultados são seguidos de uma discussão com implicações práticas e recomendações para estudos futuros.
Palavras-chave: Modelo de resposta ao item não paramétrica; Modelagem Bayesiana; Simulação Monte Carlo.

Modelos De Respuesta al Item No-Paramétricos: Comparando la Recuperación de Puntuaciones Verdaderas

Resumen
Se utilizan procedimientos no paramétricos para agregar flexibilidad a los modelos. Se propusieron tres modelos de respuesta al ítem no paramétricos, pero no se compararon directamente: Kernel smoothing (KS-IRT); la curva davidiana (DC-IRT); y el modelo bayesiano de Rasch semiparamétrico (SP-Rasch). El objetivo principal del presente estudio es comparar el desempeño de estos procedimientos en la recuperación de puntajes verdaderos simulados, utilizando puntajes de suma como puntos de referencia. El objetivo secundario es comparar su desempeño en términos de equivalencia práctica con datos reales. En general, los resultados muestran que, a excepción de DC-IRT, que es el modelo con peor desempeño, todos los otros modelos presentan resultados bastante similares a los obtenidos cuando se utilizan sumatorios. Estos resultados son seguidos por una discusión con implicaciones prácticas y recomendaciones para estudios futuros.
Palabras clave: Modelo de respuesta de ítem no paramétrico; Modelado bayesiano; simulación del Monte Carlo.

Nonparametric item response models (NIRMs) have been proposed as alternatives to traditional parametric item response theory (IRT) models. NIRMs can relax any of the pragmatic assumptions (i.e., normal density and logistic link function; Franco et al., 2022) without affecting the general assumptions (i.e., unidimensionality, latent monotonicity and local independence) of IRT. Several different procedures can be considered as NIRMs, such as Mokken scale analysis (Sijtsma & van der Ark, 2017); Bayesian semiparametric models (Falk & Cai, 2016) and nonparametric models (Karabatsos & Sheu, 2004); kernel based (Ramsay, 1991) and spline based (Ramsay & Wiberg, 2017) smooth estimate for item response functions (IRFs);

http://dx.doi.org/10.1590/1413-82712023280403

and others. However, several of these procedures have never been directly compared on their effectiveness in recovering true scores.

Nonparametric regression (Barlow & Brunk, 1972), nonparametric density estimation (Izenman, 1991) and the general nonparametric Bayesian procedure known as Dirichlet Process (DP; Ferguson, 1973) are some procedures that could be used to fit NIRMs. In general, these procedures aim at adding flexibility to the estimates, reducing underfitting while also compensating for overfitting of the data (Müller et al., 2015; Tsybakov, 2009). Applying these procedures to IRT promotes at least two advantages. First, the possibility of estimating latent IRFs that are more flexible than any parametric model. Secondly, nonparametric density estimation can aid in the estimation of true scores when skewed or multimodal distributions can be expected.

Three NIRMs are evaluated in the present study. The first is Kernel smoothing IRT (KS-IRT; Ramsay, 1991), which uses B-spline basis to estimate non-decreasing monotonic item characteristic curves (ICCs), but assumes a parametric form for the true scores. Another procedure is the Davidian curve IRT (DC-IRT; Woods & Lin, 2009), which uses B-spline basis to smooth the distribution of the estimated true scores, but assumes a parametric IRF. Finally, the semiparametric Bayesian Rasch model (SP-Rasch; San Martín et al., 2011) uses a Dirichlet Process mixture that originates smooth estimates for both ICCs and true scores. The main reason for choosing these specific procedures is due to previous results showing that they can outperform other nonparametric (Lee, 2007; Woods & Lin, 2009) and parametric (Duncan & MacEachern, 2008; Miyazaki & Hoshino, 2009) procedures on recovering true scores. Still, these models have not been compared previously. Each procedure relaxes different assumptions (i.e., parametric IRF or normal distribution for the true scores) about the item response process, so comparing them can help achieve further understanding about which or even if any parametric assumption improves the true scores' estimation.

The primary purpose of this study is to compare the three described NIRMs for dichotomous data on several simulated conditions for their effectiveness in recovering latent true scores, defined in terms of their bias, magnitude of average residuals and correlations with the known simulated true scores. The second purpose is to compare their practical equivalence (i.e., if conclusions from real data can change) using real data from a college admissions test. The rest of this paper is structured as follows. In the next three sections we present the three procedures—KS-IRT, DC-IRT, and SP-Rasch, respectively. In the fifth section, simulated and real data are used to compare the procedures. The paper ends with a discussion and some concluding remarks regarding mainly practical implications.

Kernel smoothing IRT (KS-IRT)

The basic idea of the KS-IRT (Douglas 1997; Ramsay, 1991) is to obtain a nonparametric estimate of the ICCs by taking a (local) weighted average (i.e., curve smoothing) of the probability of response Pr(θ) of the N respondents' x_{ij} responses to item i:

$$\Pr_i(\theta) = \sum_{j=1}^{N} w_j(\theta - \theta_j)x_{ij}, \quad (1)$$

where θ_j is the true score of the respondent j, θ is the vector of true scores estimates for the respondents, and the weights $w_j(\theta - \theta_j)$ are defined by a kernel function $K(\cdot)$ which imposes three conditions to these weights. They must be nonnegative; they must reach their maximum when $\theta = \theta_j$; and they will approach or equal zero as $|\theta - \theta_j|$ increases. To assure interpretability of the weights, two extra conditions are to be met: $w_j(\theta - \theta_j \geq 0)$ and $\sum_j w_j(\theta - \theta_j = 1)$. The normalizing function proposed by Nadaraya (1964) and Watson (1964) are a commonly used alternative to assure these two extra conditions:

$$w_j(\theta - \theta_j) = \frac{K\left(\frac{\theta-\theta_j}{b}\right)}{\sum_j K\left(\frac{\theta-\theta_j}{b}\right)}, \quad (2)$$

where b is the bandwidth parameter, which determines the degree of smoothing. When b is small, the bias (i.e., underfitting) is also small, but variance (i.e., overfitting) is larger.

Another requirement for using the KS-IRT is—due to the fact that θ_j is not observable—to use an estimate for the true score, denoted $\hat{\theta}_j$. A default procedure (Ramsay, 1991) is to rank the sum scores and define $F(\theta)$ as the normal cumulative distribution function (CDF), which finally leads to

$$\Pr_i(\hat{\theta}) = \frac{\sum_{j=1}^{N} K\left(\frac{\theta-\theta_j}{b}\right)x_{ij}}{\sum_{j=1}^{N} K\left(\frac{\theta-\theta_j}{b}\right)}. \quad (3)$$

The ICCs estimated by the KS-IRT procedure will also be monotonically related to the latent scores, but with

no particular parametric function defined. Some ICCs are more similar to a logistic IRF, but quite different curves can be estimated. The densities, on the other hand, will always conform to the defined CDF, which can be the CDF of any continuous distribution.

For properly fitting a KS-IRT, an adequate level of the bandwidth parameter, *b*, must be chosen. The selection of the bandwidth is a complex procedure and demands an optimal choice criterion. For selecting the bandwidth, one can use a cross-validation procedure (Wong, 1983) with the form

$$\mathrm{CV}(b_i) = \frac{1}{N}\sum_{j=1}^{N}(x_{ij}-\mathrm{Pr}_i^{-j}(\hat{\theta}_j))^{\intercal}(x_{ij}-\mathrm{Pr}_i^{-j}(\hat{\theta}_j)), \quad (4)$$

where Pr_i^{-j} is the vector of estimated probabilities of response after removing $\hat{\theta}_j$. This procedure relays on minimization algorithms, being the value of b_i that minimizes CV considered the best one for avoiding under or overfitting.

Davidian curve IRT (DC-IRT)

The Davidian curve IRT (DC-IRT) was proposed by Woods and Lin (2009) as a procedure for fitting item response models in which the distribution of the random latent variable is estimated simultaneously with the item parameters. The DC-IRT combines traditional logistic IRF models, like the two-parameter logistic (2PL) model, with a density estimation method known as Davidian curves (DCs; Zhang & Davidian, 2001). DCs are defined by:

$$DC(\theta) = P_k^2(\theta)\varphi(\theta), \quad (5)$$

where P_k is a polynomial of order k and $\varphi(\theta)$ is the standard normal density function. P_k is defined as:

$$P_k^2(\theta) = \left\{\sum_{\lambda=0}^{k} m_\lambda \theta^\lambda\right\}^2, \quad (6)$$

where λ is a nonnegative integer and m_λ are the weight coefficients for the θ^λ polynomial transformations. Two other constraints are also necessary. First, Woods and Lin (2009) used a polar coordinate transformation which assures that corrections in the estimation steps have identical density as the originally estimated θ. This assures that $DC(\theta)$, which is a probability density function, integrates to 1. The second constraint is to fix the mean and standard deviation of the initial latent density to 0 and 1, respectively, so the model is identifiable.

For effective implementation of this procedure, Woods and Lin (2009) propose an expectation maximization (EM) algorithm with two iteration steps. In the first step (E-step) random initial values for the latent scores and the item parameters are used to estimate the number of people expected to give a specific response to each item. In the second step (M-step) parameters of the DC are estimated by maximizing the following likelihood

$$L_g = \prod_{q=1}^{Q} DC(\theta_q)^{\mathrm{N}(\theta_q)} \quad (7)$$

where q is each specific quadrature point, Q is the total quadrature points, $\mathrm{N}(\theta_q)$ is the number of people expected to give a specific response on the quadrature q, and $DC(\theta_q)$ is the DC estimate for the quadrature q.

As DC-IRT models with different k parameters should be compared to achieve the best model, model selection is carried out using the Hannan–Quinn (HQ) criterion (Hannan, 1987). The HQ criterion is similar to the Akaike and to the Bayesian information criterions, although defined slightly differently

$$\mathrm{HC} = -2logLik + 2p(\log(\log(N))) \quad (8)$$

where p is the number of estimated parameters. Woods and Lin (2009) use this criterion as it has shown good performance with DCs (Davidian & Gallant, 1993; Zhang & Davidian, 2001) and Ramsay curves IRT (RC-IRT; Woods, 2006, 2007).

Semiparametric Bayesian Rasch model

In Bayesian semiparametric and nonparametric model, it is common for researchers to use the Dirichlet Process (DP; Ferguson, 1973). DP is defined as a probability distribution which the domain is a set of probability distributions, with parameters M_0 and α, representing the base distribution and a scaling parameter, respectively. The base distribution parameter can be interpreted as the expected value of the process, while the scale parameter is how similar is the realizations to the base distribution; in the limit $\alpha \to \infty$, the realization follows the same distribution as the base distribution.

The DP is typically used for estimating non regular densities (Müller & Quintana, 2004). However, it is also possible to use DPs to sample from functions (Müller et al., 2015), resulting in estimates of nonlinear and nonparametric functions. San Martín et al. (2011) proposed to use the DP to give flexibility to the parametric Rasch model. Let μ represent means, σ^2 represent variances, the estimates of true scores are θ, and the estimates of difficulties are δ; San Martín et al. (2011) model is then represented as

$$
\begin{aligned}
\delta_i &\sim Normal(\mu_\delta, \sigma_\delta^2) \\
\alpha_j &\sim Gamma(.001, .001) \\
M_{0j} &\sim Normal(\mu_\theta, \sigma_\theta^2)T(-3, 3) \\
M_j &\sim DP(\alpha_j, M_{0j}) \\
\theta_j &\sim M_j \\
\eta_{ij} &= \theta_j - \delta_i \\
\Pr_{ij} &= \frac{\exp(\eta_{ij})}{1 + \exp(\eta_{ij})} \\
X_{ij} &\sim Bernoulli(\Pr_{ij}),
\end{aligned}
\tag{9}
$$

where ~ is read as "is distributed as", and *T* represents a truncation.

The semiparametric Bayesian Rasch **(**SP-Rasch) model can be labelled as a hierarchical mixed model, which complexity implies that the posterior distribution of the parameters cannot be found in closed form (Kruschke, 2015, chapter 6). A common alternative Bayesian procedure for fitting this type of model is through the Markov Chain Monte Carlo (MCMC; Kruschke, 2015, chapter 7) method. MCMC samples estimates of the posterior distribution (which is not directly accessible) from the unnormalized posterior. Due to the descriptive and sometimes exploratory characteristic of measurement models, non-informative priors (Kruschke, 2015, chapter 10) for the parameters are usually preferred (e.g., Duncan & MacEachern, 2008). Following these procedures, irregular densities and nonparametric ICCs can be properly estimated.

Procedures' performance and hypotheses

Previous studies (Guo & Sinharay, 2011; Lee, 2007) found that KS-IRT is one of the best procedures based on smoothing of the ICCs for recovering both items and respondents' latent variables values. The formal structure of this procedure was also extended for applications in nonparametric test equating (De Ayala et al., 2018), and item selection for computerized adaptive testing (Xu & Douglas, 2006). As stated before, one of its limitations is the need to define a parametric form to $F(\theta)$. We are not aware of any study which compared the KS-IRT with other procedures that use more flexible distributions for the estimated true scores.

DC-IRT was proposed as a more efficient version of the Ramsay curve IRT (RC-IRT; Woods & Thissen, 2006), which uses B-spline-based densities estimates for the latent scores. Woods and Lin (2009) present at least two advantages of DC-IRT over RC-IRT. The first is the number of tuning parameters. RC-IRT uses three tuning parameters (the number of knots, the order of the splines, and the standard deviation of the prior distribution), DC-IRT uses only one (k). In this case, needing fewer tuning parameters is an advantage because one optimal model can be selected amongst 10 models for DC-IRT, while for RC-IRT it is one over 25 possible models. The second advantage is that DC-IRT will perform at least as well as RC-IRT in several conditions, but better when the true distribution of latent scores is skewed. This is particularly interesting as the RC-IRT has been previously shown to perform better—in a number of conditions—than parametric and other nonparametric procedures (Woods, 2006, 2007).

The SP-Rasch has not yet been directly compared to other models. Even so, similar models have shown good performance when compared to parametric models (e.g., Duncan & MacEachern, 2008). An extension of the SP-Rasch using a three-parameter logistic instead of a Rasch model as the base distribution has also shown good performance when compared to parametric models (Duncan & MacEachern, 2013). Another extension, proposed by Arenson and Karabatsos (2018), which uses no specific parametric function as base distribution, was able to perform better than the parametric 2PL model, especially when symmetric priors, as the ones used in the present study, are used. Finally, these models can all be compared to sum scores, which can be considered as lower bound benchmarks for the performance of the models (Wiberg, et al., 2018).

Simulation study

Method

Following Woods and Lin (2009), we used the 2PL model as the true IRF for data generation. Random draws of 1,000 simulated respondents and 25 items were iterated 800 times. Discriminations were drawn from a truncated normal distribution with mean equals to 1.7, standard deviation equals to .8 and bottom truncation equals to .5. Difficulties were drawn from a normal distribution with mean equals to 0 and standard deviation equals to 1.2.

Our three conditions were based on the densities used to generate the latent true scores, with values set as suggested by Woods and Lin (2009): a standard normal distribution; a skewed distribution; and a bimodal distribution. Woods and Lin (2009) used a

mixture of normal distributions for both the skewed and the bimodal distributions, where the one used for the bimodal distribution was unequal in the mixtures' standard deviations. To better represent a bimodal and a skewed distribution, we used different, but equivalent, procedures for simulating both conditions. For the skewed distribution, we used Azzalini's (1985) skewed normal distribution with parameters equal to: skewness = 1.57, mean = 0, and standard deviation = 1. For the bimodal distribution, we used a mixture of normal distributions, but with equal standard deviations and with means equally distant from the standard mean: N(-1,.5) + N(1,.5). The three conditions are presented in Figure 1.

To assess effectiveness of each procedure, three measures related to the accuracy of the estimates and one measure related to distributional properties were used. The first measure related to accuracy was Spearman's correlation, which was used to assess accuracy in ranking simulated respondents. Bias was used to assess accuracy on the range of true scores' estimates, as measured by the residuals of an additive regression between estimated and true scores. The final measure of accuracy was mean absolute error (MAE). MAE was preferred over the more common root mean square error (RMSE) because the latter increases with the variance of the distribution of error magnitudes (Willmott & Matsuura, 2005). MAE is defined as

$$\mathrm{MAE} = \frac{1}{N}\sum_{j=1}^{N}\left|\theta_j - \hat{\theta}_j\right| \tag{10}$$

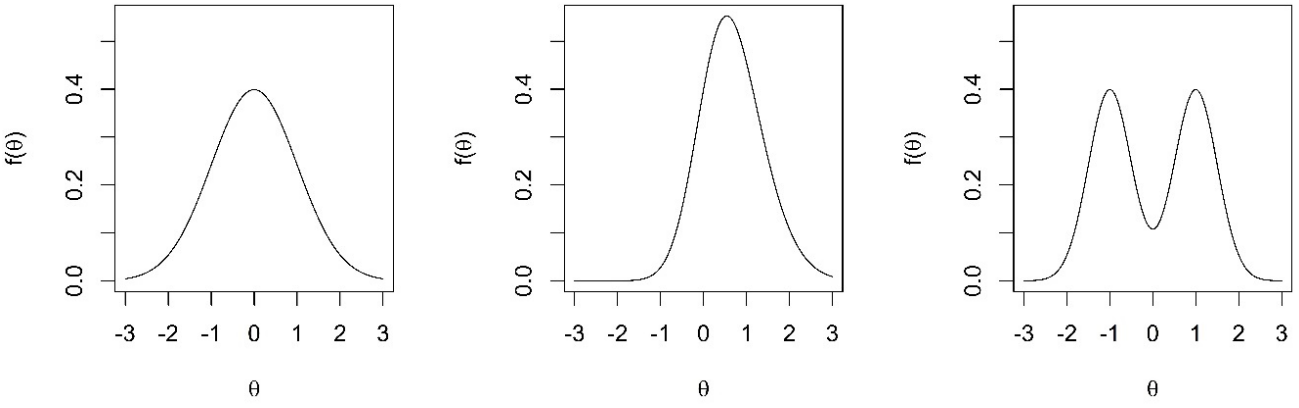
The $\hat{\theta}_j$ estimated by DC-IRT and KS-IRT were recovered using the expected a posteriori (EAP) procedure. For the SP-Rasch, true scores were recovered using the maximum a posteriori (MAP) estimate.

Distributional properties were measured using the integrated square error (ISE; Shirahata & Chu, 1992). The ISE is defined as

$$\mathrm{ISE}(\hat{g}) = \int \{\hat{g}(\theta) - g(\theta)\}^2 d\theta \tag{11}$$

where $\hat{g}(\theta)$ is the feature scaled density of the distribution of estimates of the true score θ and $g(\theta)$ is the feature scaled density of the real distribution of true scores. Both $\hat{g}(\theta)$ and $g(\theta)$ were calculated using kernel density estimates with the number of equally spaced points equal to the sample size and the bandwidth chosen adaptively using Sheather and Jones (1991) method.

All simulations and data analyses were done in R (R Core Team, 2019). We used the DC-IRT implemented in the mirt package (Chalmers, 2012). For the kernel smoothing IRT, we used the implementation in the KernSmoothIRT package (Mazza et al., 2012). For the SP-Rasch model, we used the implementation in the DPpackage package (Jara et al., 2011). The skewed distribution was generated using the rsn function from the sn package (Azzalini, 2018). The residuals of the additive regression for bias were calculated using the gam function from the mgcv package (Wood, 2012). The ISE was calculated using the density integrate.xy function from the sfsmisc package (Meachler, 2018). The code with the full simulation is available at https://osf.io/3ryz2/.



Note. From left to right: the standard normal distribution; the skewed distribution; and the bimodal distribution.

Figure 1. The three real data distributions

Results

Table 1 presents the overall results and also the findings for each condition. Numbers in bold show, for a particular condition and performance's measure, which procedure performs best. In general, sum scores, the scores estimated with KS-IRT and with the SP-Rasch are very similar in their performances. The DC-IRT procedure performed the worst in every measure and every condition. The sum scores performed the best in all conditions in terms of MAE. It also performed better in terms of ISE in the bimodal condition. The KS-IRT scores performed the best in all but the bimodal condition in terms of ISE. The SP-Rasch scores performed the best in all but the normal condition in terms of correlation's measures.

These results can be complemented with Figure 2, which includes the estimates of bias averaged over all conditions and for each condition. It is evident that all the procedures have similar bias throughout the range of possible values for the true scores throughout all the condition. However, when scores are extremely low, the SP-Rasch will be more biased than the other procedures, especially in the normal and skew conditions. Also, all the procedures will underestimate scores below the average true score, and overestimate scores above the average true score. This result was somewhat expected because of two reasons. First because all procedures estimate θ from monotonic transformations of sum scores. Second because the simulated true scores are denser around 0. Finally, it is also evident that scores below average are more biased than scores above average. This is a likely consequence from the fact that one of the true distributions is skewed, having a higher number of scores between 0 and 1 than between 0 and –1.

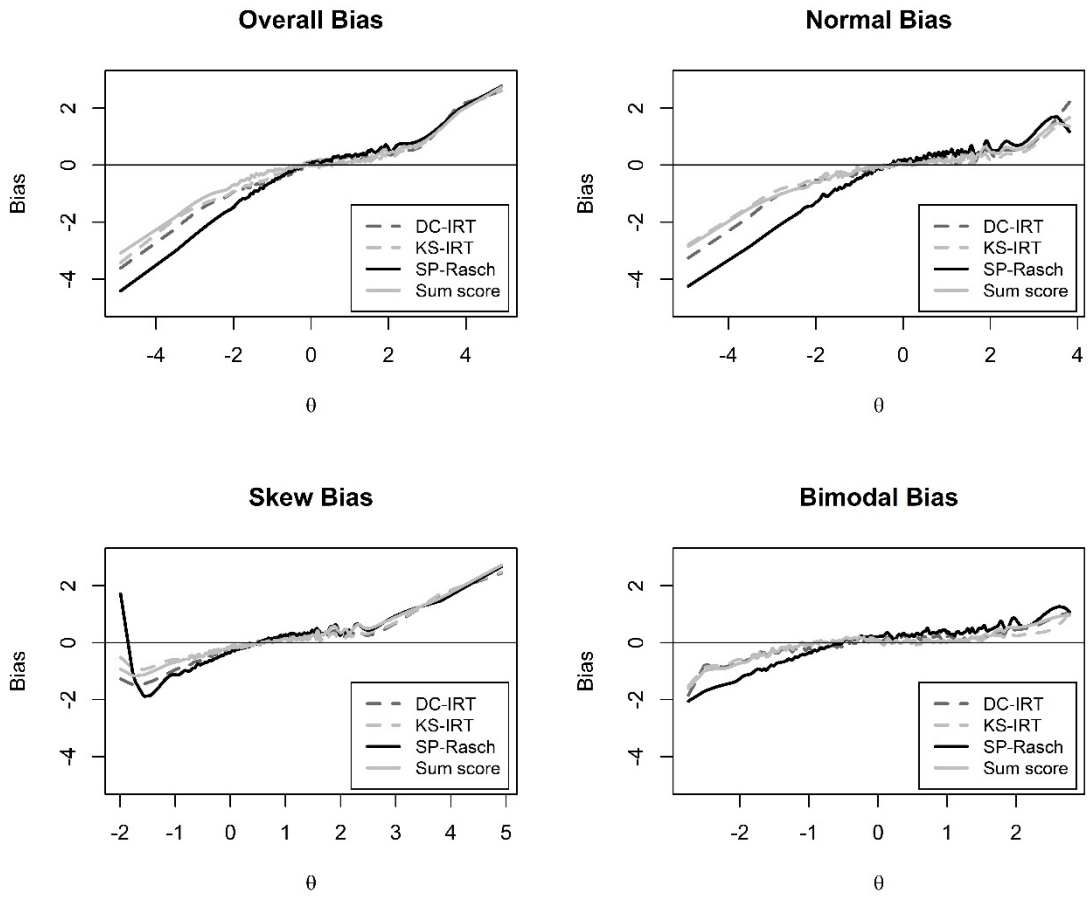
Empirical Example

For an illustrated comparison between the procedures, we used a sample of 5,000 individuals from one administration of the Swedish Scholastic Assessment

Table 1.
Average Accuracy and Distributional Properties Estimated

Condition	Procedure	Correlations	MAE	ISE
Overall	Sum score	.904	**.253**	.831
	DC-IRT	.846	.301	1.478
	KS-IRT	.903	.257	**.350**
	SP-Rasch	**.905**	.282	.811
Normal	Sum score	**.916**	**.259**	.887
	DC-IRT	.860	.303	1.201
	KS-IRT	.915	.264	**.101**
	SP-Rasch	.913	.291	.778
Skewed	Sum score	.868	**.241**	.916
	DC-IRT	.788	.285	1.159
	KS-IRT	.866	.245	**.220**
	SP-Rasch	**.869**	.264	.638
Bimodal	Sum score	.927	**.259**	**.688**
	DC-IRT	.891	.315	1.083
	KS-IRT	.928	.261	.729
	SP-Rasch	**.933**	.291	1.017

Note. DC-IRT = Davidian curve model. KS-IRT = Kernel smoothing model. SP-Rasch = Bayesian Semiparametric Rasch model. MAE = Mean Absolute Error. ISE = Integrated squared error.



Note. DC-IRT = Davidian curve model. KS-IRT = Kernel smoothing model. SP-Rasch = Bayesian Semiparametric Rasch model.

Figure 2. The bias for each procedure at estimating the true score, averaged over all conditions and specific for each condition.

Test (SweSAT). The SweSAT is a college admissions test that is administered twice a year and the scores are valid to use when applying to university for five years. It consists of a verbal and a quantitative subtest with each containing 80 items. Each subtest is scored, analyzed, and equated separately. In this example, we used the quantitative subtest of the SweSAT. We choose the quantitative test scores because they are more skewed than the verbal sum scores. This condition was chosen as it should maximize the difference of performance between the tested procedures.

The total sum score is equated and transformed to a normed score, which is used in selection to higher education in Sweden (Stage, 2003). Therefore, differences in the order, or ranking, of the respondents can result in different people accessing higher education. The estimated nonparametric scores were compared to the sum scores using three different measures. First, the densities of the nonparametric estimates were compared to the sum scores' density using the ISE. Next, Kolmogorov-Smirnov *d* statistic was calculated for the distributions of scores estimated using sum scores and the NIRMs. The *d* statistic simply represents the largest distance (in absolute value) between the CDFs of the target distribution (i.e., a normal distribution) and the distribution of the estimated scores. Finally, the Spearman's correlation was used to compare how similar the scores rank respondents, using the whole sample and the top 5% and 1% performers on the sum scores of the quantitative subtest.

Results

From Table 2 it is evident that, by evaluating the *d* statistic, the KS-IRT estimates are more normally

distributed than the other estimates. The SP-Rasch distribution was shown, by means of ISE, to be the most similar to the sum scores' distribution.

The values of *d* and ISE can also be reflected by the densities represented in Figure 3. The SP-Rasch has almost the same distribution as the DC-IRT. KS-IRT has a normal distribution and the sum scores have a positive–asymmetric distribution, with a different peak than SP-Rasch and DC-IRT.

The Spearman correlations in Figure 4 show how similarly the different procedures rank the respondents. When the whole sample is used, all procedures gave almost the same ranking of participants, with the smaller Spearman correlation equals to .98. The rank correlation between KS-IRT, SP-Rasch, and the sum scores are almost the same when we compare the procedures using only the top 5% and 1% performers. The DC-IRT presents a decrease in its rank correlation with the other procedures. Summing up, using any procedure but the DC-IRT will give basically the same rank of top performers.

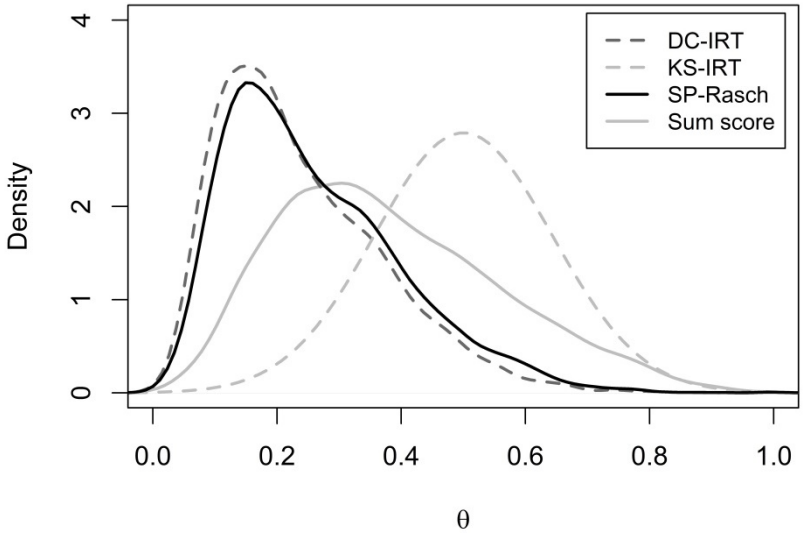
Discussion

The main aim of the present study was to compare the performance of DC-IRT, KS-IRT and SP-Rasch in recovering simulated true scores, using sum scores as benchmarks. The secondary aim was to compare their performances in terms of practical equivalence with real data. Our results support the claim that the estimated distribution of θ is arbitrary (Ramsay, 1991) and, therefore, nonparametric density estimation adds little to the effectiveness of IRMs. This follows from the fact that using an arbitrary parametric distribution for the true scores and a nonparametric procedure for ICC estimation resulted in better estimates. Using parametric IRFs and nonparametric estimation of the density of the true score, by means of DC-IRT, had

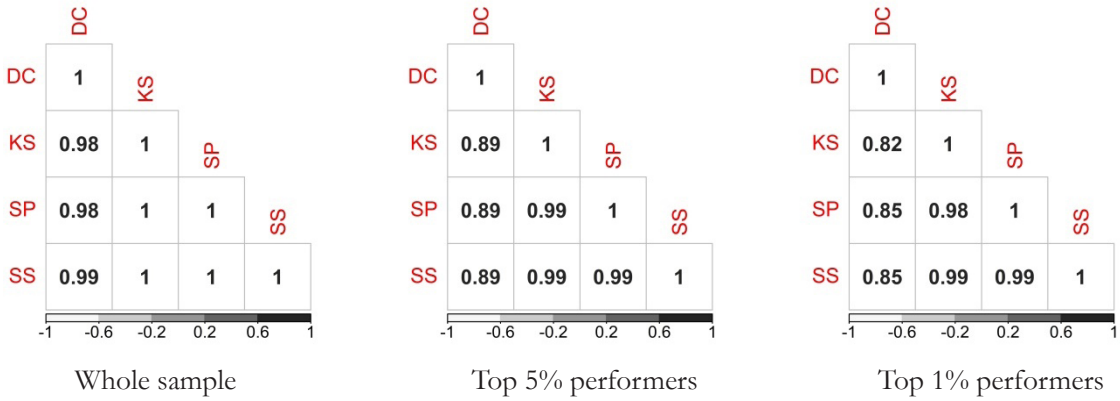
Table 2.
Distributional Properties of the Estimated Scores

Measure	Sum score	DC-IRT	KS-IRT	SP-Rasch
d	.9999	.0847	**.0002**	.310
ISE	—	9.605	8.624	**7.319**

Note. Distance to a Normal Distribution (d) and Difference from the Sum Score's Distribution (ISE). DC-IRT = Davidian curve model. KS--IRT = Kernel smoothing model. SP-Rasch = Bayesian Semiparametric Rasch model.



Note. DC-IRT = Davidian curve model. KS-IRT = Kernel smoothing model. SP-Rasch = Bayesian Semiparametric Rasch model.
Figure 3. Densities of the estimated scores



Note. DC = Davidian curve model. KS = Kernel smoothing model. SP = Bayesian Semiparametric Rasch model. SS = Sum score.

Figure 4. Correlation between scores given the whole sample, the top 1% and the top 5% performers

the least accurate performance. As expected, the KS-IRT and the SP-Rasch outperformed the sum score estimates, due to their flexibility in fitting noisy data. For ranking participants, in terms of estimates of true scores, sum scores, KS-IRT, and SP-Rasch are equally effective procedures. These results seem to support the view that the parametric assumptions about the IRFs are less relevant, or even more restricting, for estimating true scores.

One of the limitations of the present study is the fact that the non-normal distribution's conditions seemed to not cause much differences in the performances as expected. Despite following what was recommended by Woods and Lin (2009) for setting these conditions, our results show that more expressive deviations from normality are probably necessary for observing relevant differences in performance. Future studies could, for instance, test the effect of the presence of outliers (e.g., using the *t* distribution), the effect of bounded scales for the latent variables (e.g., using the beta distribution), or even a mixture of these alternatives (e.g., with mixture distributions for the latent scores).

In terms of practical implications, our results suggest that even if there are some deviations from parametric models of the true data generating process, the KS-IRT and the SP-Rasch are likely to perform similarly. This is particularly true if the aim is to rank respondents by means of estimates of true scores. Our empirical example showed that rank correlations between KS-IRT, SP-Rasch and the sum scores are very high for top performers. However, future studies, focusing on applications of these models where extreme scores on both ends of the scales are relevant, should compare how similar the methods are in ranking individuals with lower scores.

In terms of implementations of NIRMs, because the KS-IRT can be implemented using the CDF of any continuous distribution, future studies could test how changing this aspect of the model could impact the performance. Also, other techniques and methods of Bayesian modeling can be used for further extend the SP-Rasch model. For instance, differently from MLE procedures, Bayesian IRMs do not require the specification of a distribution for the sample scores, but only for the prior of possible scores (Fox, 2010). This naturally results in a more flexible distribution for the estimates of the sample true scores, somewhat similar to the DC-IRT. This can be used to extend the SP-Rasch so it relies less on both the empirical distribution of scores and the strong imposition of having a parametric, Rasch, item response function (Wiberg et al, 2018).

NIRMs have outperformed parametric IRMs in a set of different conditions on previous studies. Our research is one of few (e.g., Lee, 2007; Woods & Lin, 2009) to compare the effectiveness of different NIRMs in recovering true scores and the first, that we are aware of, to compare the SP-Rasch with other nonparametric procedures. From a theoretical point of view, our findings support that IRMs can be more efficient on recovering true scores if their ICCs are nonparametric, but show little to no improvement in allowing for asymmetric distributions on the estimates of the sample true scores. This means that we agree with Wiberg et al. (2019) that the logistic assumption of IRFs can actually

be a limitation for psychometric models which requires further understanding and changes in current research and testing practices.

References

Arenson, E., & Karabatsos, G. (2018). A Bayesian Beta-Mixture Model for Nonparametric IRT (BBM-IRT). *Journal of Modern Applied Statistical Methods, 17*(1), 1-17. https://ssrn.com/abstract=3102461

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics, 12*, 171-178. https://www.jstor.org/stable/4615982

Azzalini, A. (2018). sn: The Skew-Normal and Related Distributions Such as the Skew-t R package retrieved from *https://cran.r-project.org/web/packages/sn/index.html*.

Barlow, R. E., & Brunk, H. D. (1972). The isotonic regression problem and its dual. *Journal of the American Statistical Association*, *67*(337), 140-147. https://doi.org/10.1080/01621459.1972.10481216

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1-29. https://doi.org/10.18637/jss.v048.i06

Davidian, M., &Gallant, A. R. (1993). The nonlinear mixed effects model with a smooth random effects density. *Biometrika, 80*, 475-488. https://doi.org/10.1093/biomet/80.3.475

De Ayala, R. J., Smith, B., & Norman Dvorak, R. (2018). A comparative evaluation of kernel equating and test characteristic curve equating. *Applied Psychological Measurement*, *42*(2), 155-168. https://doi.org/10.1177/0146621617712245

Douglas, J. (1997). Joint consistency of nonparametric item characteristic curve and ability estimation. *Psychometrika*, *62*(1), 7-28. https://doi.org/10.1007/BF02294778

Duncan K. A., MacEachern S. N. (2013). Nonparametric Bayesian modeling for item response with a three parameter logistic prior mean. In Edwards M. C., MacCallum R. C. (Eds.), *Current topics in the theory and application of latent variable methods* (pp. 108-125). New York, NY: Routledge

Duncan, K. A., & MacEachern, S. N. (2008). Nonparametric Bayesian modelling for item response. *Statistical Modelling*, *8*(1), 41-66. https://doi.org/10.1177/1471082X0700800104

Falk, C. F., & Cai, L. (2016). Semiparametric item response functions in the context of guessing. *Journal of Educational Measurement*, *53*(2), 229-247. https://doi.org/10.1111/jedm.12111

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics, 1*(2), 209-230. https://www.jstor.org/stable/2958008

Fox, J. P. (2010). *Bayesian item response modeling: Theory and applications*. New York: Springer.

Franco, V. R., Laros, J. A., Wiberg, M., & Bastos, R. V. S. (2022). How to think straight about psychometrics: Improving measurement by identifying its assumptions. *Trends in Psychology*, 1-21. https://doi.org/10.1007/s43076-022-00183-6

Guo, H., & Sinharay, S. (2011). Nonparametric item response curve estimation with correction for measurement error. *Journal of Educational and Behavioral Statistics*, *36*(6), 755-778. https://doi.org/10.3102/1076998610396891

Hannan, E. J. (1987). Rational transfer function approximation. *Statistical Science, 2*, 135-161. https://www.jstor.org/stable/2245658

Izenman, A. J. (1991). Review papers: Recent developments in nonparametric density estimation. *Journal of the American Statistical Association*, *86*(413), 205-224. https://doi.org/10.1080/01621459.1991.10475021

Jara, A., Hanson, T. E., Quintana, F. A., Müller, P., & Rosner, G. L. (2011). DPpackage: Bayesian semi- and nonparametric modeling in R. *Journal of Statistical Software*, *40*(5), 1-30.

Karabatsos, G., & Sheu, C. F. (2004). Order-constrained Bayes inference for dichotomous models of unidimensional nonparametric IRT. *Applied Psychological Measurement*, *28*(2), 110-125. https://doi.org/10.1177/0146621603260678

Kruschke, J. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Cambridge: Academic Press.

Lee, Y. S. (2007). A comparison of methods for nonparametric estimation of item

characteristic curves for binary items. *Applied Psychological Measurement*, *31*(2), 121-134. https://doi.org/10.1177/0146621606290248

Mazza, A., Punzo, A., & McGuire, B. (2012). KernSmoothIRT: An R package for kernel smoothing in item response theory. *arXiv preprint arXiv:1211.1183*.

Meachler, M. (2018). sfsmisc: Utilities from 'Seminar fuer Statistik' ETH Zurich. R Package retrieved from *https://cran.r-project.org/web/packages/sfsmisc/index.html*.

Miyazaki, K., & Hoshino, T. (2009). A Bayesian semiparametric item response model with Dirichlet process priors. *Psychometrika*, *74*(3), 375-393. https://doi.org/10.1007/s11336-008-9108-6

Müller, P., & Quintana, F. A. (2004). Nonparametric Bayesian data analysis. *Statistical Science, 19*(1), 95-110. https://doi.org/10.1214/088342304000000017

Müller, P., Quintana, F. A., Jara, A., & Hanson, T. (2015). *Bayesian nonparametric data analysis*. New York: Springer.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, *9*(1), 141-142. https://doi.org/10.1137/1109020

R Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*(4), 611-630. https://doi.org/10.1007/BF02294494

San Martín, E., Jara, A., Rolin, J. M., & Mouchart, M. (2011). On the Bayesian nonparametric generalization of IRT-type models. *Psychometrika*, *76*(3), 385-409.

Sheather, S. J., & Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, *53*(3), 683-690. https://doi.org/10.1111/j.2517-6161.1991.tb01857.x

Shirahata, S., & Chu, I. S. (1992). Integrated squared error of kernel-type estimator of distribution function. *Annals of the Institute of Statistical Mathematics*, *44*(3), 579-591.

Sijtsma, K., & van der Ark, L. A. (2017). A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. *British Journal of Mathematical and Statistical Psychology*, *70*(1), 137-158. https://doi.org/10.1111/bmsp.12078

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. New York: Springer.

van der Linden, W.J., & Hambleton, R.K. (Eds.) (2013). *Handbook of modern item response theory*. New York: Springer.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 359-372. https://www.jstor.org/stable/25049340

Wiberg, M., Ramsay, J. O., & Li, J. (2018). Optimal scores: An alternative to parametric item response theory and sum scores. *Psychometrika*, 1-13. https://doi.org/10.1007/s11336-018-9639-4

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*(1), 79-82.

Wong, W. H. (1983). On the consistency of cross-validation in kernel nonparametric regression. *The Annals of Statistics*, *11*(4), 1136-1141. https://www.jstor.org/stable/2241303

Wood, R. (1978). Fitting the Rasch model: A heady tale. *British Journal of Mathematical and Statistical Psychology, 31*, 27–32. https://doi.org/10.1111/j.2044-8317.1978.tb00569.x

Wood, S. (2012). mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation. R Package retrieved from *https://cran.r-project.org/web/packages/mgcv/index.html*.

Woods, C. M. (2006). Ramsay-curve item response theory to detect and correct for nonnormal latent variables. *Psychological Methods, 11*, 253-270. https://doi.org/10.1037/1082-989X.11.3.253

Woods, C. M. (2007). Ramsay-curve IRT for Likert type data. *Applied Psychological Measurement, 31*, 195-212. https://doi.org/10.1177/0146621606291567

Woods, C. M., & Lin, N. (2009). Item response theory with estimation of the latent density using Davidian curves. *Applied Psychological Measurement*, *33*(2), 102-117. https://doi.org/10.1177/0146621608319512

Woods, C. M., & Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika*, *71*(2), 281-301. https://doi.org/10.1007/s11336-004-1175-8

Xu, X., & Douglas, J. (2006). Computerized adaptive testing under nonparametric IRT models. *Psychometrika*, *71*(1), 121-137. https://doi.org/10.1007/s11336-003-1154-5

Zhang, D., & Davidian, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics*, *57*(3), 795-802. https://doi.org/10.1111/j.0006-341X.2001.00795.x

Recebido em: 09/07/2021
Reformulado em: 24/03/2023
Aprovado em: 26/05/2023

About the authors:

Vithor Rosa Franco is a Professor at São Francisco University, Psychology Department, Campus Campinas, R. Waldemar César da Silveira, 105, Jardim Cura D'Ars (SWIFT), SP - Brazil. His research interests include measurement theory and quantitative modeling, being especially interested in Bayesian and computational methods for psychological research.
ORCID: https://orcid.org/0000-0002-8929-3238
E-mail: vithorfranco@gmail.com

Marie Wiberg is a full professor in Statistics (with specialty in psychometrics) at the Department of Statistics of the Umeå School of Business, Economics and Statistics, Umeå University, SE-90187, Umeå, Sweden. Her research interests include educational measurement and psychometrics in general. Special interest is given to test equating, parametric and nonparametric item response theory, and international large-scale assessments.
ORCID: https://orcid.org/0000-0001-5549-8262
E-mail: marie.wiberg@umu.se

Rafael Valdece Sousa Bastos is a master's student at São Francisco University, University, Psychology Department, Campus Campinas, R. Waldemar César da Silveira, 105, Jardim Cura D'Ars (SWIFT), SP - Brazil. His research interests include measurement theory and quantitative psychology, being especially interested in comparison of methods used in psychological research.
ORCID: https://orcid.org/0000-0003-2444-6982
E-mail: rafavsbastos@gmail.com

Contact:

Vithor Rosa Franco
São Francisco University, Psychology Department, Campus Campinas
Rua Waldemar César da Silveira, 105, Jardim Cura D'Ars (SWIFT)
São Paulo-SP, Brasil