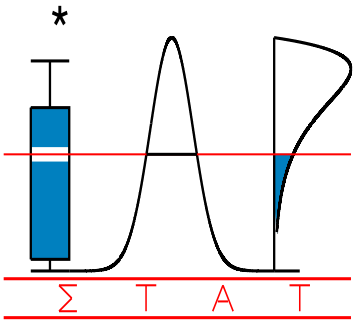
T E C H N I C A L

R E P O R T

08029

ON THE ANALYSIS OF BAYESIAN SEMIPARAMETRIC IRT-TYPE MODELS

SAN MARTIN, E., JARA, A., ROLIN, J.-M. and M. MOUCHART



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

On the Analysis of Bayesian Semiparametric IRT-type Models

ERNESTO SAN MARTÍN[1,2]*, ALEJANDRO JARA[3],

JEAN-MARIE ROLIN[4], AND MICHEL MOUCHART[4]

[1] *Departamento de Estadística, Pontificia Universidad Católica de Chile, Santiago, Chile*

[2] *Measurement Center MIDE UC, Pontificia Universidad Católica de Chile, Santiago, Chile*

[3] *Biostatistical Centre, Department of Public Health, Catholic University of Leuven, Belgium*

[4] *Institut de statistique, Université catholique de Louvain, Louvain-la-Neuve, Belgium*

April 14, 2008

Abstract

Motivated by the characteristics of two educational datasets, we study the Bayesian identification and consistency of semiparametric IRT-type models, where the uncertainty on the abilities' distribution is modeled using a prior distribution on the space of probability measures. We establish sufficient conditions for the identification and consistency in the Bernoulli and Poisson versions of the Rasch model. For unbounded count (resp. binary) responses the parameters are identified when a finite (resp. infinite) number of probes

*Author for correspondence: Ernesto San Martín, Department of Statistics and Measurement Center MIDE UC, Pontificia Universidad Católica de Chile, Santiago, Chile. E-mail : esanmart@mat.puc.cl

are available and they are consistently estimated when the number of subjects tends (resp. subjects and probes tend) to infinite. The validity of the findings as potential necessary conditions are evaluated using simulated data.

Keywords: Bayesian identification, Bayesian consistency, Rasch model, Rasch Poisson Count Model, Dirichlet processes, Polya Tree processes

1 Introduction

Item Response Theory (IRT) models are widely used in educational measurement (see e.g., De Boeck and Wilson, 2004, and references therein). Rasch-type models (Rasch, 1960) are typical examples of this class and can be viewed as a particular case of Generalized Linear Models (GLM). In the Rasch-type models, the linear predictor η_{ij} depends on two parameters in an additive way $\eta_{ij} = \theta_i - \beta_j$, where $\theta_i \in \mathbb{R}$ corresponds to the ability of subject i, and $\beta_j \in \mathbb{R}$ corresponds to the difficulty of probe/item j.

From a modeling point of view, these models are intended to analyze individual differences. Because of that, the attention is focused on the ability parameter θ_i, considered as an incidental parameter (see, Rasch, 1960; Andersen, 1980). Therefore, Rasch-type models are an example of the Neyman-Scott phenomenon (Neyman and Scott, 1948), namely, that the maximum likelihood estimator of the difficulty parameters is inconsistent due to the presence of the incidental parameters (see e.g., Andersen, 1980; Ghosh, 1995). By assuming that the ability parameters are independent random variables with a common probability distribution G, Kiefer and Wolfowitz (1956) and Pfanzagl (1970; 1993) proved the classical consistency of both the difficulty parameters $\boldsymbol{\beta}_{1:n} = (\beta_1, \ldots, \beta_n)^T$ and the distribution G, provided their identification. Note that in this case, the GLM representation of Rasch-type models are considered as conditional mod-

els and the model is completed by specifying a marginal latent model describing the process generating the abilities, G. In this way, Rasch-type models become a member of the Generalized Linear Mixed Model (GLMM) class (see e.g., De Boeck and Wilson, 2004), where the difficulty and ability parameters are interpreted as fixed and random effects, respectively.

Typically, the probability distribution G is assumed to be a member of a parametrized family $\{G_{\boldsymbol{\varphi}} : \boldsymbol{\varphi} \in \boldsymbol{\Theta} \subset \mathbb{R}^p\}$, commonly a normal distribution, with unknown parameter vector $\boldsymbol{\varphi}$. In such a case, $(\boldsymbol{\beta}_{1:n}, \boldsymbol{\varphi})$ are estimated using available software packages such as SAS, Stata or BUGS. However, in applications there are almost inevitable concerns about the lack of robustness of resulting inferences with respect to assumed forms of distributional components (see e.g., Agresti et al., 2004; Heagerty and Kurland, 2001; Verbeke and Lesaffre, 1996). Furthermore, the process generating the individual abilities is viewed as the responsible of the generation of the observed phenomenon (Lord, 1969; Bartholomew, 1987; Borsboom et al., 2003). Therefore, the assumption of a specific parametric distribution may be too restrictive to represent the actual between-subject variation. This leads to consider both the items/probes difficulties and a general probability distribution G as the parameters of interest.

In this paper we consider Bayesian semiparametric inference which replaces the traditional normal assumption with Bayesian nonparametric models. Dirichlet process (DP) (Ferguson, 1973), Mixtures of DP (MDP) (Antoniak, 1974), Polya trees (PT) (Ferguson, 1974), and Mixtures of PT (MPT) (Lavine, 1992) have been used to define flexible nonparametric models for random effects distributions in the GLMM context. We refer to Bush and MacEachern (1996), Mukhopadhyay and Gelfand (1997) and Kleinman and Ibrahim (1998a, 1998b), for implementations based on DP and MDP priors, and Walker and Mallick (1997), Hanson (2006), and Jara et al. (2007) for implementations based on PT and MPT priors.

We argue that it is important to ask whether a Bayesian semiparametric formulation of Rasch-type models has an empirical sense. In other words, since the statistical inference is only related to the characteristics of the distribution of the observed variables, it is of interest to know whether different $(\boldsymbol{\beta}_{1:n}, G)$ can be distinguished from the observations. Thus, it is important to consider the corresponding identification problem. It must be stressed that the identification problem we are dealing with is more general than the one of identifying the latent distribution G in a mixture model. The latter problem has been considered in Teicher (1961) and Chandra (1977). On the other hand, in a classical context, De Leeuw and Verhelst (1986) proved the consistency of $\boldsymbol{\beta}_{1:n}$ and G, when G is specified as a step-function probability distribution (see also Lindsay et al., 1991). Therefore, this result cannot be used to imply identification for a general G. Moreover, the authors explicitly leave open the case of a general probability distribution (see De Leeuw and Verhelst, 1986, pp. 192-193). Overcoming these problems largely motivates the developments presented in this paper.

It must be pointed out that identification issues present no formal difficulties to a Bayesian analysis. Indeed, the remark of Lindley (1971, p. 46), "In passing it might be noted that unidentifiability causes no real difficulty in the Bayesian approach", recognizes the fact that a proper prior is transformed into a proper posterior using the sampling model and the probability calculus. However, if the interest is not related to predictions but focuses on inference about a non identified parameter, then such formal assurances have little practical value. In fact identification problems describe some aspects of the Bayesian learning process from the data. For instance, as more data become available, the posterior mass will not concentrate on a point in the model, making asymptotic analysis difficult. Identification problems may also induce computational problems as they could imply ill-behaved posterior surfaces and Markov chain Monte Carlo (MCMC) methods can be difficult to implement in such cases.

The motivation for this work comes from two education measurement applications. In the first case, we consider data from the SIMCE Chilean national project. This project has developed mandatory tests to assess regularly the educational progress at three levels: 4th and 8th grades in primary school, and 2th grade in secondary school. We focus on data from the Math test, applied in 2006 to 243,834 second grader examinees in secondary school. The test consists of 45 multiple choice items with 4 alternatives, including a variety of questions ranging from problem formulation, functions, simple algebra, geometry and probability. The response $Y_{ij} \in \{0, 1\}$ is a binary variable indicating whether the individual i answers in a correct manner the item j. In the second case, we consider data from the Ministry of Education of the Flemish government (Belgium). They developed a test to evaluate the curriculum for French as a foreign language for pupils of the first stage of secondary school. Specifically, a written test was developed in which the students had to copy six different texts of different lengths. We consider data from a calibration study where 483 pupils were evaluated. The response variable of interest was the number of spelling errors in each test $Y_{ij} \in \mathbb{A} \subset \mathbb{N}$. Standard analysis for this type of data consider a Poisson conditional model viewed as a limiting case of binomial trials, $Y_{ij} \in \mathbb{N}$ (Rasch, 1960; Jansen and van Dujin, 1992; Jansen, 1994).

Two characteristics of our motivating examples are of interest for the present work. In the SIMCE dataset, the data has a finite and small support points, but a large number of items were considered. On the other hand, in the French Written dataset the data points can be considered as realizations of random variables with an infinity support, but a small number of probes were considered. In this paper we study whether this two extreme and different contexts provide enough information to identify the model parameters $(\boldsymbol{\beta}_{1:n}, G)$. Specifically, we establish sufficient conditions for the identification and consistency of the Bayesian semiparametric binary Rasch model and Rasch Poisson Count model.

The rest of the paper is organized as follows. In Section 2, we begin with a discussion of the identification concept, with an emphasis on the differences between the classical and Bayesian versions. Although most of the material in this section is not original, the discussion is necessary for the development of the semiparametric Rasch models. In particular, we build on the concept of parameter minimal sufficiency. Section 3 develops the framework of the Bayesian semiparametric modeling. In Sections 4 and 5 we provide sufficient conditions for identification and consistency, respectively. In Section 6, we present simulated data examples and the analysis of our motivating examples. The simulated datasets allow the comparison of various Bayesian nonparametric approaches and the evaluation of the implications of the sufficient identification restrictions. We conclude with a short discussion in Section 7.

2 The Identification Concept

Whereas the classical or sampling approach considers a statistical model as an indexed family of distributions on the sample space, the Bayesian approach considers a unique probability measure on the product space "parameters $\times$ observations". This produces two different approaches to the identification, namely the injectivity of a mapping in a sampling theory approach and the minimal sufficiency of the parameterization in a Bayesian approach. These approaches are discussed next. We refer to the Section 4.6.2 in (Florens et al., 1990), for the technical conditions required for linking these approaches.

2.1 Classical Identification

A classical statistical model is defined as a family of sampling probability distributions indexed by a parameter,

$$\left\{(\Omega, \mathcal{X}), P^{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\right\} \tag{1}$$

where $(\Omega, \mathcal{X})$ is the sample space, $P^{\boldsymbol{\theta}}$ is a sampling probability on $(\Omega, \mathcal{X})$ indexed by a parameter $\boldsymbol{\theta}$, and $\boldsymbol{\Theta}$ is the parameter space. Note that expression (1) can be viewed either as a family of probability distributions explaining a population of infinite size from which the data are randomly drawn (in the point of view of Fischer, 1922), or as a model implied by a structural formulation (in the point of view of Koopmans and Reiersøl, 1950). The identification of any statistical model deals with the identification of a parametrization.

Definition 1 *A parametrization $\boldsymbol{\theta}$ is said to be globally identified by the data if the mapping $\boldsymbol{\theta} \longmapsto P^{\boldsymbol{\theta}}$ is injective.*

From the definition, when a parametrization $\boldsymbol{\theta}$ is not identified, the parameter space $\boldsymbol{\Theta}$ can be decomposed into a partition of subsets of the form $[\boldsymbol{\theta}] = \{\boldsymbol{\theta}^{'} \in \boldsymbol{\Theta} : P^{\boldsymbol{\theta}} = P^{\boldsymbol{\theta}^{'}}\}$. The parameters belonging to $[\boldsymbol{\theta}]$ are said to be observationally equivalent. The identification of $\boldsymbol{\theta}$ is intended to avoid this parameter redundancy. An identified parametrization generates, therefore, a partition into singletons of the parameter space. In the context of parametric models, the parametric space is simply a subset of $\mathbb{R}^p$, were p is the dimension of the real vector $\boldsymbol{\theta}$. In these models, Rothenberg (1971) introduce the concept of local identification which occurs when there may be a number of observationally equivalent structures but they are isolated from each other.

Definition 2 *A parameter point $\boldsymbol{\theta}$ is said to be locally identified if there exists an open neighborhood of $\boldsymbol{\theta}$ containing no other $\boldsymbol{\theta}^{'} \in \boldsymbol{\Theta}$ which is observationally equivalent.*

It should be noted that global identification implies local identification (Rothenberg, 1971) and that local identification everywhere in the parameter space $\boldsymbol{\Theta}$ is a necessary but not sufficient condition for global identification (see e.g., Bechger et al., 2001). Note also that local identification at a point of $\boldsymbol{\Theta}$ does not imply local identification everywhere in $\boldsymbol{\Theta}$. Rothenberg (1971) showed that, under weak regularity conditions, local identification is equivalent to non-singularity of the theoretical or observed information matrix.

2.2 Bayesian Identification

Bayesian identification is defined in terms of parameter minimal sufficiency. This formal analogy was first noted by Kadane (1974) and Picci (1977). We next discuss the concept of sufficiency and minimal sufficiency at the parameter level, first introduced in Barankin (1961) and Barankin et al. (1980).

2.2.1 Sufficient Parameter

A natural way to introduce the definition of sufficient parameter is by taking advantage of the symmetric role between parameters and observations in a Bayesian context (see Mouchart and Rolin, 1984; Florens et al., 1990). Thus, the definition of sufficient parameter is parallel to the definition of sufficient statistic by replacing "observations" by "parameters". In a sampling context, a statistic $S(\boldsymbol{Y})$ is a sufficient statistic for $\boldsymbol{\theta}$ if the conditional distribution of the sample $\boldsymbol{Y}$ given $S(\boldsymbol{Y})$ does not depend on $\boldsymbol{\theta}$. When $\boldsymbol{\theta}$ is a random variable such a definition is equivalent to $\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{\theta} \mid S(\boldsymbol{Y})$.

Definition 3 *A function $g(\boldsymbol{\theta})$ of the parameter $\boldsymbol{\theta}$ is a sufficient parameter for $\boldsymbol{Y}$ if the conditional distribution of the sample $\boldsymbol{Y}$ given $\boldsymbol{\theta}$ is the same that the distribution of the sample $\boldsymbol{Y}$ given $g(\boldsymbol{\theta})$, that is,*

$$\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{\theta} \mid g(\boldsymbol{\theta}). \tag{2}$$

Condition (2) implies that the distribution of $\boldsymbol{Y}$ is completely determined by $g(\boldsymbol{\theta})$, or in other words, $\boldsymbol{\theta}$ is redundant once $g(\boldsymbol{\theta})$ is known. By the symmetry of a conditional independence relation, it can also be concluded that $g(\boldsymbol{\theta})$ is a sufficient parameter if the conditional distribution of the redundant part $\boldsymbol{\theta}$ given the sufficient parameter $g(\boldsymbol{\theta})$, it is not updated by the sample, i.e., $p(\boldsymbol{\theta} \mid \boldsymbol{Y}, g(\boldsymbol{\theta})) = p(\boldsymbol{\theta} \mid g(\boldsymbol{\theta}))$ (Dawid, 1979; Florens et al., 1990).

2.2.2 Minimal Sufficiency and Bayesian Identification

In a Bayesian model defined by the joint distribution of $(\boldsymbol{Y}, \boldsymbol{\theta})$, the basic question of parametric sufficiency is whether the parameter (of interest) $\boldsymbol{\theta}$ provides the minimal description of the sampling process $(\boldsymbol{Y} \mid \boldsymbol{\theta})$. In other words, whether the parameter (of interest) $\boldsymbol{\theta}$ is a minimal sufficient one, that is, if $\boldsymbol{\theta}$ is a sufficient parameter and it is a function of any other sufficient parameter. This is precisely the problem of Bayesian identification (see Florens and Rolin, 1984; Florens and Mouchart, 1986).

Definition 4 *Consider the Bayesian model defined by the joint probability distribution on* $(\boldsymbol{Y}, \boldsymbol{\theta})$. *The parameter* $\boldsymbol{\theta}$ *is said to be Bayesian identified or b-identified, if it is a minimal sufficient parameter.*

In words, a parameter is said to be b-identified if it corresponds to the greatest possible parameter reduction for which the prior information is updated by the sample. Consequently, a b-identified parameter fully characterizes the learning process underlying a Bayesian model.

Heuristically, a minimal sufficient parameter is generated by the family of all the sampling expectations of (integrable) functions f defined on the sample space, namely $E[f(\boldsymbol{Y}) \mid \boldsymbol{\theta}]$. The parameter (of interest) $\boldsymbol{\theta}$ need not be equal to the minimal sufficient parameter; when the equality holds, $\boldsymbol{\theta}$ is b-identified. This equality can precisely be formulated in σ-algebraic terms. The identified σ-field (on the parameter space) may be constructed through the σ-algebraic projection of the observation on the parameter space, *i.e.* the σ-field on Θ generated by the sampling expectation of all integrable function defined on the sample space $(\Omega, \mathcal{X})$. When the identified σ-field coincides with the the σ-field of subsets of Θ, the parameter (of interest) become b-identified; for details, see Florens et al. (1990). A formal characterization of the minimal sufficient parameter is given in Section A of the Appendix.

3 Bayesian Semiparametric Rasch-type models

Model identification strongly depends on the model specification. It is relevant, therefore, to make explicit the hypotheses underlying semiparametric Rasch models. In this section, we develop these aspects in two steps. We first deal with general stochastic dependencies between observable variables, latent variables, and parameters, which are described by means of conditional independence assumptions. In a second step, we specify the probability distributions compatible with the underlying dependencies structures.

3.1 The General Class of Rasch Models

Assume that for each of m subjects the responses to n probes $\{Y_{ij}, i = 1, \ldots, m, j = 1, \ldots, n\}$ are recorded. Let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in})'$ be the response pattern for subject i, where $Y_{ij} \in V \subseteq \mathbb{N}$. Let $\{\theta_i : i = 1, \ldots, m\}$ and $\{\beta_j : j = 1, \ldots, n\}$ be sets of real-valued random variables where θ_i represents the *ability* of subject i and β_j characterizes a property of the item j (typically called *difficulty parameter*). Let G and H_n be two probability measures on $(\mathbb{R}, \mathcal{B})$ and $(\mathbb{R}^n, \mathcal{B}^n)$, representing the distribution of the θ_i's and of $\boldsymbol{\beta}_{1:n} = (\beta_1, \cdots, \beta_n)$, respectively. For all $m, n \in \mathbb{N}$, denote $\boldsymbol{Y}_{1:m} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m)$ and $\boldsymbol{\theta}_{1:m} = (\theta_1, \ldots, \theta_m)$. The general class of Rasch models is specified through the following structural properties:

H1. For all m and n, $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m \perp\!\!\!\perp G, H_n \mid \boldsymbol{\theta}_{1:m}, \boldsymbol{\beta}_{1:n}$.

H2. The process generating $(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m \mid \boldsymbol{\theta}_{1:m}, \boldsymbol{\beta}_{1:n})$ is characterized by the following properties:

(i) $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m$ are mutually independent given $(\boldsymbol{\theta}_{1:m}, \boldsymbol{\beta}_{1:n})$, i.e., $\perp\!\!\!\perp_{1 \le i \le m} \boldsymbol{Y}_i \mid \boldsymbol{\theta}_{1:m}, \boldsymbol{\beta}_{1:n}$.

(ii) For all i, $\boldsymbol{Y}_i \perp\!\!\!\perp \boldsymbol{\theta}_{1:m}, \boldsymbol{\beta}_{1:n} \mid \theta_i, \boldsymbol{\beta}_{1:n}$.

(iii) For all i, $\perp\!\!\!\perp_{1 \le j \le n} Y_{ij} \mid \theta_i, \boldsymbol{\beta}_{1:n}$ (axiom of local independence).

(iv) For all i, $Y_{ij} \perp\!\!\!\perp \boldsymbol{\beta}_{1:n} \mid \theta_i, \beta_j$ for all $1 \leq j \leq n$.

H3. The process generating $\boldsymbol{\theta}_{1:m}, \boldsymbol{\beta}_{1:n} \mid G, H_n$ is characterized by the following properties:

(i) $\boldsymbol{\theta}_{1:m} \perp\!\!\!\perp \boldsymbol{\beta}_{1:n} \mid H_n, G$.

(ii) $\boldsymbol{\theta}_{1:m} \perp\!\!\!\perp H_n \mid G$.

(iii) $\boldsymbol{\beta}_{1:n} \perp\!\!\!\perp G \mid H_n$ and $\boldsymbol{\beta}_{1:n} \mid H_n \sim H_n$.

(iv) $\underset{1\leq i\leq m}{\perp\!\!\!\perp} \theta_i \mid G$ and $\theta_i \mid G \sim G$ for all i.

H4. $G \perp\!\!\!\perp H_n$.

Different prior distributions can be specified on G. This is discussed in Section 3.2.2. The probability distribution H_n remains unspecified. It will have a role only later on when discussing the identification of the semiparametric Rasch model. The structural properties H1 - H4, implies an important characteristic of Rach-type models, namely, the observations $\{\boldsymbol{Y}_i : 1 \leq i \leq m\}$ form an *iid* process conditionally on $(\boldsymbol{\beta}_{1:n}, G)$. The common distribution of $\boldsymbol{Y}_i$, for all $i \in \mathbb{N}$, is given by

$$\begin{aligned} P[\boldsymbol{Y}_i = \boldsymbol{y}_i \mid \boldsymbol{\beta}_{1:n}, G] &= E\left\{P[\boldsymbol{Y}_i = \boldsymbol{y}_i \mid \theta_i, \boldsymbol{\beta}_{1:n}] \mid \boldsymbol{\beta}_{1:n}, G\right\}, \\ &= \int \left\{ \prod_{1\leq j\leq n} P[Y_{ij} = y_{ij} \mid \theta, \beta_j] \right\} G(d\theta), \end{aligned} \tag{3}$$

for $\boldsymbol{y}_i \in V^n$. In this model we consider $\boldsymbol{\beta}_{1:n}$ and G as the parameters of interest. It is important to stress that the *iid* property of the statistical model is implied by the structural hypotheses of the model rather than assumed. This property may be lost if some of these hypotheses are changed.

3.2 Distributional Structure

The semiparametric model specification is completed by choosing the specific parametric forms of $(Y_{ij} \mid \theta_i, \beta_j)$ and the prior distribution for G.

3.2.1 The Parametric Component

We concentrate the attention on two parametric specifications. Specifically, we consider the Rasch Poisson Count model (RPCM) and the binary Rasch model (RM) (Rasch, 1960). In the first case, the sampling distribution is given by,

$$(Y_{ij} \mid \theta_i, \beta_j) \sim \text{Poisson}(\exp(\theta_i - \beta_j)), \tag{4}$$

where Y_{ij} is an unbounded count variable, typically representing the number of miss-reading/miss-copying for the subject i in the text j. In the RM, Y_{ij} is a binary coding the correct answer of individual i to the item j. In this case, it is assumed that

$$(Y_{ij} \mid \theta_i, \beta_j) \sim \text{Bernoulli}\left(\Psi(\theta_i - \beta_j)\right), \tag{5}$$

where $\Psi(x) = \exp(x)/(1 + \exp(x))$.

3.2.2 The Bayesian Nonparametric Component

The nonparametric component of model (3) corresponds to the probability distribution $G \in \mathcal{P}(\mathbb{R}, \mathcal{B})$ generating the individual abilities, where $\mathcal{P}(\mathbb{R}, \mathcal{B})$ denotes the set of the probability measures defined on the Borel space $(\mathbb{R}, \mathcal{B})$. Accordingly, the uncertainty about the distribution G is characterized by a nonparametric prior distribution. Here we focus on DP (Ferguson, 1973), MDP (Antoniak, 1974), PT (Ferguson, 1974), and MPT (Lavine, 1992) priors because they robustify parametric models by embedding them in a larger encompassing nonparametric model. In our context, DP and PT priors are centered on a normal $N(\mu, \sigma^2)$ distribution while MDP and MTP are center on the family of normal distributions by placing a prior distribution on μ and σ^2. In order to fix the notation, we introduce the nonparametric models next.

3.2.3 Dirichlet Processes and Variations

The DP was introduced by Ferguson (1973) as a prior probability distribution on the space of probability measures defined, for instance, on $(\mathbb{R}, \mathcal{B})$.

Definition 5 *A random probability measure G follows a Dirichlet process with parameters (M, G_0), where $M \in \mathbb{R}_0^+$ and $G_0 \in \mathcal{P}(\mathbb{R}, \mathcal{B})$, written as $G \mid M, G_0 \sim DP(MG_0)$, if for any measurable nontrivial partition $\{B_l : 1 \leq l \leq k\}$ of $\mathcal{B}$, $\{G(B_l) : 1 \leq l \leq k\}$ has a Dirichlet distribution with parameters $(MG_0(B_1), \ldots, MG_0(B_k))$.*

Some useful properties of the DP are that $G(\mathbb{R}) = 1$ a.s. and that $G(B)$ ~Beta$(MG_0(B)$,$MG_0(\overline{B}))$. Therefore, $E[G(B)] = G_0(B)$ and $Var[G(B)] = G_0(B)G_0(\overline{B})/(M+1)$. These results show the role of G_0 and M, namely, that G is centered around G_0 and that M is a precision parameter.

The analysis under a DP prior distribution requires the specification of the base measure $\alpha = MG_0$. Antoniak (1974) considers mixtures of Dirichlet processes (MDP) where the parameter α of the DP is allowed to be random. By so doing, a hierarchical model of the following type is defined: $u \mid P \sim P$ and $G \mid u \sim DP(\alpha_u)$. We used the notation

$$G \mid P \sim \int DP(\alpha_u) P(du). \tag{6}$$

In practice, one may propose a parametric family as the base measure and put hyper-priors on the parameters of that family and/or to allow the precision parameter M to be random. In the applications considered in Section 6, α_u is defined by a normal distribution such that $u = (\mu, \sigma^2, M)$.

Although DP and MDP are simple and computationally tractable priors for unknown distributions, they produce distributions that are discrete with probability one, making it unsuitable

for density modeling. This can be avoided by convolving the random distribution distribution with some continuous kernel, or more generally, by using a DP or MDP to define a mixture distribution with infinitely many components of some simple parametric form (see e.g., Ferguson, 1983; Lo, 1984; Escobar and West, 1995). An alternative approach are the Polya Trees processes (PT), which under some sufficient conditions have, with probability one, trajectories on the set of continuous or absolutely continuous probability distributions.

3.2.4 Polya Trees Priors and Variations

PT priors were originally considered by Ferguson (1974) and latter studied thoroughly by Mauldin et al. (1992) and Lavine (1992; 1994). To define the PT processes, the following notation is needed. Let $E = \{0, 1\}$, $E^0 = \emptyset$, E^l be the l-fold product $E \times \cdots \times E$ and $E^* = \bigcup_{m=0}^{\infty} E^m$. Let $\Pi = \{\pi_l \, : \, l \in \mathbb{N}\}$ be a separating binary tree of partitions of $\mathbb{R}$; that is, let $\pi_0, \pi_1, \ldots$ be a sequence of partitions such that $\bigcup_{l=0}^{\infty} \pi_l$ generates the Borel σ-field and such that every $B \in \pi_{l+1}$ is obtained by splitting some $B' \in \pi_l$ into two pieces. Thus, each partition π_l has 2^l elements that can be represented in the form $\{B_\epsilon \, : \, \epsilon \in E^l\}$, with $B_\emptyset = \mathbb{R}$ and, for all $\epsilon = \epsilon_1 \cdots \epsilon_l \in E^*$, $B_{\epsilon 0}$ and $B_{\epsilon 1}$ correspond to the two pieces into which B_ϵ is split.

Definition 6 *Let Π be a separating binary tree of partitions and let $\mathcal{A} = \{\alpha_\epsilon \in \mathbb{R}^+ \, : \, \epsilon \in E^*\}$. A random probability measure G on $(\mathbb{R}, \mathcal{B})$ is said to have a Polya tree distribution with parameters $(\Pi, \mathcal{A})$, written $G \sim PT\,(\Pi, \mathcal{A})$, if there exist random variables $\mathcal{Y} = \{Y_\epsilon : \epsilon \in E^*\}$ such that the following hold:*

(i) The collection $\mathcal{Y}$ consists of mutually independent random variables.

(ii) For every $\epsilon \in E^*$, $Y_{\epsilon 0} \sim$ Beta $(\alpha_{\epsilon 0}, \alpha_{\epsilon 1})$.

(iii) For every $l = 1, 2, \ldots$ and every $\epsilon \in E^l$,

$$G(B_{\epsilon_1 \cdots \epsilon_l}) = \left(\prod_{j=1, \epsilon_j = 0}^{l} Y_{\epsilon_1 \ldots \epsilon_{j-1}} \right) \left(\prod_{j=1, \epsilon_j = 1}^{l} (1 - Y_{\epsilon_1 \ldots \epsilon_{j-1}}) \right),$$

where the first term in the product is interpreted as $Y_\emptyset$ *or* $1 - Y_\emptyset$.

In practical applications, it would be difficult to specify the family $\mathcal{A}$ and the partition Π. This leads to consider a special class of PT, namely, a PT centered around a given probability distribution G_0 defined on $(\mathbb{R}, \mathcal{B})$. More specifically, let $G_0(t)$ be the cdf of G_0. To center the PT around G_0, we let each level l of the partition $\Pi = \{\pi_l \ : \ l \in \mathbb{N}\}$ to coincide with the intervals $(G_0^{-1}\{k/2^l\}, G_0^{-1}\{(k+1)/2^l\}]$ for $k = 0, 1, \ldots, 2^l - 1$, and further taking $\alpha_{\epsilon 0} = \alpha_{\epsilon 1}$ for all $\epsilon \in E^*$. In this case, the (prior) expectation of G is equal to G_0. Thus, G_0 has a role similar to that of the center measure of a DP. Once the PT is centered around a probability measure G_0, the family $\mathcal{A}$ further determines how much G can "deviate" from G_0. Besides, the PT will have infinitely many more parameters which may be used to describe one's prior belief. To simplify such a specification, a default method is adopted, where α_ϵ is chosen depending on the length of the finite string ϵ only. This choice needs some careful thought, as the parameters in $\mathcal{A}$ directly controls the type of trajectories generated by the process. Ferguson (1974) pointed out that $\alpha_\epsilon = 1$ yields a random probability G that is continuous singular with probability one, whereas that $\alpha_{\epsilon 1 \cdots \epsilon l} = l^2$ yields a random probability G that is absolutely continuous w.r.t. the Lebesgue measure with probability one. In general, any function $\rho(l)$ such that $\sum_{l=1}^{\infty} \rho(l)^{-1} < \infty$ guarantees G to be absolutely continuous. An example of such a specification is the family $c\, l^2$ with $c > 0$ (Hanson and Johnson, 2002). In this paper, we consider the family $\mathcal{A}^c = \{\alpha_{\epsilon_1 \cdots \epsilon_l} = c\, l^2 \ : \ \epsilon \in E^*, \, l \in \mathbb{N}\}$ to perform the Bayesian inference of semiparametric IRT-type models.

As in the case of DP, it may be difficult in practice to specify a single centering distribution to center the PT. Moreover, once it is specified, the inference can be unduly affected. One way to

mitigate these problems is to specify a Mixture of Polya Trees (MPT) (see Lavine, 1992; Hanson and Johnson, 2002). A MPT is induced for G by considering a PT where the partitioning set Π is indexed by a specific parameter ζ. This additional parameter will average out jumps to yield smooth densities. For details, see Hanson and Johnson (2002). For the MPT, the following notation is used

$$G \mid P \sim \int PT(\Pi^{\zeta}, \mathcal{A}^{c})P(d\zeta, dc). \tag{7}$$

4 Identification of Bayesian semiparametric IRT-type models

4.1 Identification of the Semiparametric Rasch Poisson Count model

In this section we derive sufficient conditions for the identification of $(\boldsymbol{\beta}_{1:n}, G)$ in the SRPCM. The basic strategy is based on the Theorem 1 (p. 151) in Mouchart and San Martin (2003). The usage of this theorem requires the concept of p-completeness of a random variable, which we introduce next.

Definition 7 *Let $(\Omega, \mathcal{M}, P)$ a probability space and, for $i = 1, 2, 3$, let X_i be a random variable defined from $(\Omega, \mathcal{M})$ to a measurable space $(S_i, \mathcal{S}_i)$. Let $p \in [1, \infty]$. X_1 is said to be p-complete w.r.t. X_2, which is denoted as $X_1 \ll_p X_2$, if for all Borel function $h : (S_1, \mathcal{S}_1) \longrightarrow (\mathbb{R}, \mathcal{B})$ such that $E(|h(X_1)|^p) < \infty$ (when $p = \infty$, it is taken the essential supremum), the following implication follows:*

$$E[h(X_1) \mid X_2] = 0 \;\; P\text{-a.s.} \Longrightarrow h(X_1) = 0 \;\; P\text{-a.s.} \tag{8}$$

Moreover, conditionally on X_3, X_1 is p-complete w.r.t X_2, which is denoted as $X_1 \ll_p X_2 \mid X_3$, if and only if $(X_1, X_3) \ll_p (X_2, X_3)$.

Since $E[\cdot \mid X_2]$ is a linear operator on the p-integrable functions, expression (8) means that the null space of $E[\cdot \mid X_2]$ reduces to $\{0\}$. This is equivalent to the injectivity of the conditional expectation $E[\cdot \mid X_2]$ (see Conway, 1985, pp. 376). In other words, the p-completeness of X_1 w.r.t. X_2 means that the expectation conditional on X_2 is an injective operator defined on $L^p(\Omega, \mathcal{M}_1, P)$. Thus, if X_1 plays the role of a "statistic" and X_2 of a "parameter", then $X_1 \ll_p X_2$ corresponds to the Bayesian counterpart of the classical definition of a complete statistic and can be viewed as the injectivity of the sampling expectation on the p-integrable functions of the statistic. Similarly, if X_1 plays the role of a "parameter" and X_2 of a "statistic", $X_1 \ll_p X_2$ corresponds to the injectivity of the posterior expectation on the p-integrable functions of the parameter; for details and properties we refer to the Chapter 5 in Florens et al. (1990). They refer to the concept of p-completeness of a parameter as "p-strong identification" due to the fact that this corresponds to a sufficient condition for b-identification. For a comparison between the classical and the Bayesian completeness, see also San Martín and Mouchart (2007).

Mouchart and San Martin (2003) use the concept of p-completeness to show that the 2-strong identification of $(\boldsymbol{\theta}_{1:m}, \boldsymbol{\beta}_{1:n})$ by $\boldsymbol{Y}_{1:m}$ in the conditional model $p(\boldsymbol{Y}_{1:m} \mid \boldsymbol{\theta}_{1:m}, \boldsymbol{\beta}_{1:n})$, and the b-identification of G by $\boldsymbol{\theta}_{1:m}$ in the marginal model $p(\boldsymbol{\theta}_{1:m} \mid G)$, imply the b-identification of the parameters $(\boldsymbol{\beta}_{1:n}, G)$ by $\boldsymbol{Y}_{1:m}$ in the statistical model $p(\boldsymbol{Y}_{1:m} \mid \boldsymbol{\beta}_{1:n}, G)$. This result is used to obtain the following theorem.

Theorem 1 *In the Bayesian semiparametric Rasch Poisson count model defined by the structural properties H1 - H4, and (4), the difficulty parameters $\boldsymbol{\beta}_{2:n}$ and the distribution G generating the individual abilities are b-identified by $\boldsymbol{Y}_1$ conditionally on β_1, provided that at least two probes are considered.*

The proof of Theorem 1, relies on the following arguments (see, Section B, Appendix, for a complete proof):

(1) If $Y \mid \lambda \sim \text{Poisson}(\lambda)$, then $\lambda \ll_2 Y$ for any prior distribution $m(\lambda)$ defined on $(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+))$.

(2) Since $\exp(\cdot)$ is a bijective function and $Y_{ij} \mid \theta_i, \beta_j \sim \text{Poisson}(\exp(\theta_i - \beta_j))$, step (1) implies that,

$$\theta_i - \beta_j \ll_2 Y_{ij} \quad \forall\, 1 \leq j \leq n. \tag{9}$$

For $n \geq 2$, using assumptions H2.ii and H2.iii, condition (9) imply that $(\theta_i - \beta_1, \ldots, \theta_i - \beta_n) \ll_2 \boldsymbol{Y}_i \mid \beta_1$, which is equivalent to,

$$\theta_i, \boldsymbol{\beta}_{2:n} \ll_2 \boldsymbol{Y}_1 \mid \beta_1. \tag{10}$$

(3) Since $\theta_i \mid G \sim G$, we have that G is b-identified by θ_i. Moreover, since $\theta_i \perp\!\!\!\perp \beta_1 \mid G$, a property implied by both H3.i and H3.ii, it follows that G is b-identified by θ_i conditionally on β_1.

Theorem 1 states the identification of $(\boldsymbol{\beta}_{2:n}, G)$ for $n \geq 2$ conditionally on β_1. In practice, this leads to fix the value of β_1, for instance, at 0. Note that the identification restrictions established in this theorem do not involve any particular form of the prior distribution H_n and of the prior distribution on G, although the prior null sets of β_1 play a important role in (10).

4.2 Identification of the semiparametric Rasch model

Because of the finite support of the Bernoulli distribution, the arguments used in the proof of Theorem 1 cannot be applied to obtain identifying restrictions for the SRM. As a matter of fact, the identification analysis of the SRPCM involves the 2-completeness of the parameter of

a Poisson distribution, which can be established thanks to its countably infinite support.

For analyzing the identification of the SRM, it is useful to distinguish between the identified parametrization of the semiparametric Rasch model and the parametrization of interest $(\boldsymbol{\beta}_{1:n}, G)$. In the case of the SRM, the identified parameters are the probabilities of the 2^n different possible patterns responses

$$\gamma_{\boldsymbol{y}_1} \doteq P[\boldsymbol{Y}_1 = \boldsymbol{y}_1 \mid \boldsymbol{\beta}_{1:n}, G] \;\; = \;\; \int_{\mathbb{R}} \prod_{j=1}^{n} \Psi(\theta - \beta_j)^{y_{1j}} \left[1 - \Psi(\theta - \beta_j)\right]^{1-y_{1j}} G(d\theta),$$

where $\boldsymbol{y}_1 \in \{0,1\}^n$ takes values among the 2^n potential individual response patterns. Therefore, the identification analysis of the SRM reduces to express the parameter of interest as functions of the identified parametrization. For a particular $\boldsymbol{y}_1$, let $J = \{j \;:\; y_{ij} = 1\}$. Using hypothesis H2.iii, $\gamma_{\boldsymbol{y}_1}$ can be written as

$$\gamma_{\boldsymbol{y}_1} = P\left[\bigcap_{j\in J}\{Y_{1j} = 1\} \cap \bigcap_{j\in J^c}\{Y_{1j} = 0\} \mid \boldsymbol{\beta}_{1:n}, G\right] =$$

$$= \int_{\mathbb{R}} \prod_{j\in J} \Psi(\theta - \beta_j) \times \prod_{j\in J^c} \{1 - \Psi(\theta - \beta_j)\}\, G(d\theta)$$

$$= \exp\left\{-\sum_{j\in J} \beta_j\right\} \times \int_{\mathbb{R}} \frac{\exp\{|J|\theta\}}{\prod\limits_{1\le j\le n} \{1 + \exp\{\theta - \beta_j\}\}}\, G(d\theta), \tag{11}$$

where $|J|$ is the cardinality of set J. Expression (11) represents $2^n - 1$ independent relations that can be used to identify the parameters in the model. By taking $J = \{j\}$ in (11) and dividing by the same relation evaluated at $J = \{1\}$, it follows that, for $2 \le j \le n$, $\delta_j \doteq \beta_j - \beta_1$ is a function of the identified parametrization of the Rasch model and, therefore, identified by $\boldsymbol{Y}_1$.

Expression (11) can be re-parameterized in terms of $\{\delta_j \;:\; j = 2, \ldots, n\}$ as follows

$$\gamma_{\boldsymbol{y}_1} = \exp\left\{-\sum_{j\in J} \delta_j\right\} \times \int_{\mathbb{R}} \frac{\exp\{|J|u\}}{\prod\limits_{1\le j\le n} \{1 + \exp\{u - \delta_j\}\}}\, G_{\beta_1}(du), \tag{12}$$

where $G_{\beta_1}((-\infty, x]) \doteq G((-\infty, x + \beta_1])$ for $x \in \mathbb{R}$. This probability transformation is well defined since, by hypotheses H3.i and H3.ii, $\theta_1 \perp\!\!\!\perp \beta_1 \mid G$. The expression $\exp(-\sum_{j \in J} \delta_j)$ is a function of $\{\delta_j : j = 2, \ldots, n\}$ and, therefore, of the identified parametrization. The integral at the right hand side of expression (12) is a function of both $\{\delta_j : j = 2, \ldots, n\}$ and G_{β_1}. Taking into account that $\{\delta_j : j = 2, \ldots, n\}$ are identified, $2^n - n - 2$ equations remains to identify some characteristic of G_{β_1}.

Theorem 2 *For the semiparametric Rasch model defined by the structural properties H1 - H4 and (5), the difficulty parameters $\boldsymbol{\beta}_{2:n}$ and $2^n - n - 2$ characteristics of G are identified by $\boldsymbol{Y}_1$ conditionally on β_1.*

Theorems 1 and 2 clearly show that the support of the data play an important role in the identification of the SRM and SRPCM. For the SRM, the quantity of characteristics of G that can actually be identified depends on the number of items, whereas for the SRPCM the complete distribution G is identified with at least two probes. It seems natural to ask wether the full latent distribution G is identified with an infinite quantity of items in the SRM. In what follows we establish sufficient conditions for the identification of the difficulty parameters and the latent distribution G in the asymptotic experiment (when $n \to \infty$).

Theorem 3 *Consider an asymptotic Bayesian semiparametric Rasch model obtained when $n \to \infty$, with the structural properties H1 - H4 and (5). The difficulty parameters $\boldsymbol{\beta}_{1:\infty}$ and the latent distribution G generating the individual abilities are b-identified by $\boldsymbol{Y}_1$ if the following two conditions hold:*

(i) The difficulty parameters $\boldsymbol{\beta}_{1:\infty}$ form an iid process with common probability distribution H.

(ii) At least one of the following identifying restrictions hold:

(ii.1) β_1 *is fixed.*

(ii.2) G *is a.s. a probability distribution on* $\mathbb{R}$ *such that* $\int xG(dx) = \mu$*, where* $\mu \in \mathbb{R}$ *is a known constant.*

(ii.3) G *is a.s. a probability distribution on* $\mathbb{R}$ *with a known* q*-quantile (*i.e. *there exists* (q, x_q) *such that* $G(x_q) = q$ *a.s.).*

As pointed out in Section 2, a minimal sufficient parameter is generated by the family of all sampling expectations of (integrable) functions defined on the sample space. In the case of Theorem 3, these expectations are of the form $E[f(\boldsymbol{Y}_1) \mid \boldsymbol{\beta}_{1:\infty}, G]$, where $\boldsymbol{Y}_1 = (Y_{ij} \ : \ 1 \leq j < \infty)$. Since these expectations are measurable functions of $(\boldsymbol{\beta}_{1:\infty}, G)$, the proof of Theorem 3 consists in establishing that $(\boldsymbol{\beta}_{1:\infty}, G)$ is a measurable function of such sampling expectations. Let us sketch the proof of Theorem 3 (see Section C of the Appendix for a complete proof):

(1) Denote by $\mathcal{A}^*$ the minimal sufficient parameter of the asymptotic experiment described by the random variables $(\boldsymbol{Y}_1, \boldsymbol{\beta}_{1:\infty}, G)$. Using the same arguments as in Theorem 2, it follows that $\boldsymbol{\delta}_{2:\infty} \doteq \{\delta_j \ : \ j \geq 2\}$ are measurable functions of $\mathcal{A}^*$.

(2) Using the fact that the difficulty parameters $\boldsymbol{\beta}_{1:\infty}$ form an *iid* process with a common probability distribution H, and using the Conditional Law of Strong Numbers and the Conditional Dominated Convergence Theorem, it follows that G_{β_1} is also a measurable function of $\mathcal{A}^*$.

(3) From steps (1) and (2) it follows that $\boldsymbol{\delta}_{2:\infty}$ and G_{β_1} are measurable w.r.t. $\mathcal{A}^*$. Since $\boldsymbol{\beta}_{1:\infty}$ and $(\boldsymbol{\delta}_{2:\infty}, \beta_1)$ are in a one-to-one relation, the theorem follows if a restriction is imposed leading to establish a one-to-one relation between G_{β_1} and (β_1, G). By so doing, $(\boldsymbol{\beta}_{1:\infty}, G)$ becomes measurable w.r.t. $\mathcal{A}^*$. There are at least two ways to establish such a relationship: (i) to establish the identification fixing β_1 (*i.e.* in the conditional experiment given β_1), or (ii) to fix either the location of G thought its mean or a q-quantile.

Theorem 3 establish the identification of the items parameters and the distribution G w.r.t. the asymptotic Bayesian experiment characterized by the random variable $(\boldsymbol{Y}_1, \boldsymbol{\beta}_{1:\infty}, G)$, jeopardizing thus the empirical meaning of G under a finite number of items.

Remark 1 Since the response patterns $\{\boldsymbol{Y}_m : m \in \mathbb{N}\}$ form an *iid* process conditionally on $(\boldsymbol{\beta}_{1:n}, G)$, for all $n \in \mathbb{N}$, the identification analysis is entirely equivalent if we consider either the infinite sequence of individual patterns $\boldsymbol{Y}_{1:\infty}$, or only one individual pattern $\boldsymbol{Y}_1$ (see Florens et al., 1990, Theorem 9.3.12). In order words, in an *iid* process the sample size has no role for the identification analysis.

5 Bayesian consistency

The main asymptotic issue in Bayesian nonparametric models is one of consistency. In a Bayesian context, the analysis of consistency may be motivated by the "what if" principle introduced by Diaconis and Freedman (1986). A prior or posterior distribution stands for the current knowledge about a parameter. A perfect knowledge implies a degenerate prior. Thus consistency means convergence, in some sense, of knowledge towards the perfect knowledge with increasing amount of data. Two styles of asymptotics could be considered. Frequentist style asymptotics studies the behavior of the posterior with respect to draws from a fixed sampling distribution. Bayesian style asymptotics studies the behavior with respect to the marginal distribution obtained by integrating the sampling distribution with respect to the prior. Here we consider the convergence of posterior expectations w.r.t. the joint probability measure defined on the product space "parameters $\times$ observations". From this perspective, the almost sure convergence is treated as a genuinely Bayesian concept for being an asymptotic feature of the joint probability on the parameter space and on the sampling space. Moreover, using Doob's Martingale Theorem, a L^1-convergence of posterior expectations can also be established, where the L^1-space is defined w.r.t. the unique probability measure defined on the product space "pa-

rameters $\times$ observations".

Let $(\boldsymbol{Y}_{1:\infty}, \boldsymbol{\vartheta})$ be defined on a common probability space $(\Omega, \mathcal{M}, P)$ and let $\boldsymbol{\phi} \doteq f(\boldsymbol{\vartheta})$ be the parameter of interest. It is well known that, by the Martingale Theorem, $E(\boldsymbol{\phi} \mid \boldsymbol{Y}_{1:m})$ *always* converges P-almost surely and in $L^1(\Omega, \mathcal{M}, P)$ to $E[\boldsymbol{\phi} \mid \boldsymbol{Y}_{1:\infty}]$, the posterior expectation of $\boldsymbol{\phi}$ given the infinite sequence of observations, provided that $\boldsymbol{\phi}$ is an integrable function. It seems natural to say that the Bayesian estimator $E[\boldsymbol{\phi} \mid \boldsymbol{Y}_{1:m}]$ is consistent if $E[\boldsymbol{\phi} \mid \boldsymbol{Y}_{1:\infty}] = \boldsymbol{\phi}$ a.s. This fact motivates the following definition given by Florens and Rolin (1984).

Definition 8 *Let $\{\boldsymbol{Y}_{1:m} : m \in \mathbb{N}\}$ and $\boldsymbol{\vartheta}$ be random variables defined on a common probability space $(\Omega, \mathcal{M}, P)$, and let $f \in L^1(\Omega, \mathcal{M}, P)$. The posterior expectation $E[f(\boldsymbol{\vartheta}) \mid \boldsymbol{Y}_{1:m}]$ is said to be a Bayesian consistent estimator for $f(\boldsymbol{\vartheta})$ if it converge P-a.s. to $f(\boldsymbol{\vartheta})$.*

Definition 8 is equivalent to say that ϕ is a.s. a function of $\boldsymbol{Y}_{1:\infty}$. This means that $Var(\mathbb{1}_A \mid \boldsymbol{Y}_{1:\infty}) = 0$ for all $A \in \sigma(\phi)$, the σ-field generated by ϕ. Therefore, Bayesian consistency means that the posterior probability of an event in $\sigma(\boldsymbol{\phi})$ is a.s. 0 or 1, and it formalizes the fact that $\boldsymbol{\phi}$ is "perfectly known" after the observation of the (infinite) sample. We refer to Sections 7.4 and 7.5 in Florens et al. (1990) for more details and for a discussion about the relationships between this concept of Bayesian consistency and the concepts defined with respect to a family of sampling distributions.

The Bayesian consistency of $(\boldsymbol{\beta}_{1:n}, G)$ for the SRM and SRPCM is based on a more general result summarized in the following theorem proved by Florens and Rolin (1984) and Florens et al. (1990).

Theorem 4 *Let $\{\boldsymbol{Y}_{1:m} : m \in \mathbb{N}\}$ and $\boldsymbol{\vartheta}$ be random variables defined on a common probability space $(\Omega, \mathcal{Y}, P)$. If $\{\boldsymbol{Y}_{1:m} : m \in \mathbb{N}\}$ form an iid process conditionally on $\boldsymbol{\vartheta}$, then the b-identified parameter $\boldsymbol{\vartheta}^*$ is consistently estimated by $E(\boldsymbol{\vartheta}^* \mid \boldsymbol{Y}_{1:m})$.*

5.1 Bayesian consistency of the Semiparametric Rasch Poisson Count model

Since the $\{\boldsymbol{Y}_m \,:\, m \in \mathbb{N}\}$ form an *iid* process given $(\boldsymbol{\beta}_{1:n}, G)$ (see Section 3.1), it follows that, conditionally on β_1, $\{\boldsymbol{Y}_m \,:\, m \in \mathbb{N}\}$ are still *iid* given $(\boldsymbol{\beta}_{2:n}, G)$. Therefore, conditionally on β_1, we have that, for all integrable measurable function h,

$$E[h(\boldsymbol{\beta}_{2:n}, G) \mid \boldsymbol{Y}_{1:m}, \beta_1] \longrightarrow h(\boldsymbol{\beta}_{2:n}, G) \qquad P^{\beta_1}\text{-a.s.},$$

since, conditionally on β_1, $(\boldsymbol{\beta}_{2:n}, G)$ is b-identified by $\boldsymbol{Y}_1$ (see Theorem 1). By taking expectation w.r.t. β_1, we obtain the following corollary.

Corollary 1 *For the semiparametric Rasch Poisson count model defined by the structural properties H1 - H4 and (4), along with the identification restriction stated in Theorem 1, we have that, for all integrable measurable function h,*

$$\lim_{m\to\infty} E[h(\boldsymbol{\beta}_{1:n}, G) \mid \boldsymbol{Y}_{1:m}] = h(\boldsymbol{\beta}_{1:n}, G) \quad P\text{-a.s.}$$

In particular, the difficulty parameters β_j's and the probability distribution G are consistently estimated by their respective posterior expectations.

5.2 Bayesian consistency of the Semiparametric Rasch model

Note that when an infinite number of items is available, the *iid* property is valid in the Rasch model, i.e., $\{\boldsymbol{Y}_m \,:\, m \in \mathbb{N}\}$ are *iid* conditionally on $(\boldsymbol{\beta}_{1:\infty}, G)$. Here $\boldsymbol{Y}_m \in \{0,1\}^\infty$ because of the infinite number of items. Now, for any finite and fixed number of items k, $\boldsymbol{\beta}_{1:k}$ is a function $\boldsymbol{\beta}_{1:\infty}$. But $(\boldsymbol{\beta}_{1:\infty}, G)$ is identified by $\boldsymbol{Y}_1$ conditionally on β_1. Therefore, by Theorem 4, $(\boldsymbol{\beta}_{1:\infty}, G)$ is a.s. a function of $(\boldsymbol{Y}_{1:\infty}, \beta_1)$. It follows that, for all k fixed, $\boldsymbol{\beta}_{1:k}$ is a.s. a function of $(\boldsymbol{Y}_{1:\infty}, \beta_1)$. Finally, by the Martingale Convergence Theorem, it follows that, for all integrable function h,

$$\lim_{m\to\infty}\lim_{n\to\infty} E[h(\beta_{1:k}, G) \mid \boldsymbol{Y}^n_{1,1}, \ldots, \boldsymbol{Y}^n_{m,1}] = h(\beta_{1:k}, G) \qquad P^{\beta_1}\text{-a.s.}$$

where $\boldsymbol{Y}_{i,1}^n = (Y_{i1}, \ldots, Y_{in}) \in \{0,1\}^n$. By taking expectation w.r.t. β_1, the following corollary is obtained.

Corollary 2 *In the Bayesian semiparametric Rasch model defined by the structural properties H1 – H4 and (5), along with the identification restrictions stated in Theorem 3, for all integrable measurable function h,*

$$\lim_{m\to\infty}\lim_{n\to\infty} E[h(\beta_{1:k}, G) \mid \boldsymbol{Y}_{1,1}^n, \ldots, \boldsymbol{Y}_{m,1}^n] = h(\beta_{1:k}, G) \qquad P\text{-a.s.}$$

In particular, the difficulty parameters $\boldsymbol{\beta}_{1:k}$ and the probability distribution G are consistently estimated by their respective posterior expectations in the double asymptotic experiment.

Remark 2 The Martingale Theorem ensures that the convergence results established in Corollaries 1 and 2 are also valid in $L^1(\Omega, \mathcal{M}, P)$, provided that h is an integrable measurable function.

6 Applications

In this section, the evaluation of the identification restrictions developed in sections 4.1 and 4.2, are explored by means of both simulated and our motivating educational datasets. Specifically, using simulated data we illustrate the behavior of the Bayes estimators under small and relatively large number of probes. The simulated datasets also allow the comparison of various Bayesian nonparametric approaches. Individual abilities are generated from a distribution much different from the usual normal parametric case. Moreover, the nonparametric prior distributions have been centered on the usual parametric specification. The first real life example focuses on data from the SIMCE project developed to assess the educational progress in Chile. The second example considers data from a writing test for the evaluation of the curriculum standards of French as a foreign language in Belgium. Functions implementing MCMC algorithms

to fit the models were written in a compiled language and incorporated into the R (R Development Core Team, 2006) library "DPpackage" (Jara, 2007). This library was used in the analyses presented here.

6.1 Empirical evaluation of the identification restrictions

Two theoretical differences between SPCRM and SRM have been established in Sections 4 and 5, namely, (i) that the parameters of interest of the SPCRM are identified when a finite number of probes is available, whereas in the SRM the identification holds only when an infinite number of items is considered; and (ii) that the parameters of interest of the SPCRM are consistently estimated in the asymptotic experiment (i.e., when the number of examinees m tends to infinite), whereas in the SRM the parameters are consistently estimated in the doubly asymptotic experiment (*i.e.*, when $m \to \infty$ and $n \to \infty$). In agreement with these results, when a test is composed of n items, Theorem 3 establishes that only $2^n - n - 2$ characteristics of G can be identified. Nevertheless, G cannot be consistently estimated with a finite n number of items. The objective of this section is to illustrate these differences by means of simulations. More precisely, we compare the Bayesian semiparametric models under three different number of items, $n = 2$, $n = 4$, and $n = 40$. The simulated data sets consists of $m = 250$ subjects and of abilities simulated from a mixture of normal distributions with two components $0.5N(-1, 0.25) + 0.5N(2, 0.125)$. The same simulated random effects were used to generated the data set for the generation of the Rasch and Rasch Poisson data. The true density and a histogram of the 250 abilities are shown in Figure 1.

[Figure 1 about here.]

We compare the models with respect to the estimation of the mixing distribution, G, by considering the Kolmogorov distance $KD = \max_y |\widehat{G}(b \mid \boldsymbol{Y}) - G(b)|$ between the true distribution

function $G(b)$ and the posterior predictive distribution function $\widehat{G}(\cdot \mid \boldsymbol{Y})$ obtained from each approach. As the PT priors were specified in order to assign the random distribution with probability one to the set of continuous distributions, the L_1 distance $L_1(\boldsymbol{Y}) = \int |\widehat{g}(b \mid \boldsymbol{Y}) - g(b)|db$ between the true density $g(b)$ and the posterior predictive density $\widehat{g}(b \mid \boldsymbol{Y})$, was also considered. Note that a more thorough comparison that accounts for sampling variability would involve calculating the mean L_1 distance $ML_1(\boldsymbol{Y}) = E[L_1(\boldsymbol{Y})]$, where expectation is with respect to the joint distribution of $\boldsymbol{Y}$, but this is beyond the scope of this illustration. Finally, regarding formal model determination, a cross validation model comparison criteria was adopted. The so-called log Pseudo Marginal Likelihood, LPML $= \sum_{i=1}^{n}\sum_{j=1}^{k} \log \text{CPO}_{ij}$, where $\text{CPO}_{ij} = p(Y_{ij} \mid \boldsymbol{Y}^{-ij})$, developed by Geisser and Eddy (1979), was considered.

When the baseline parameters were considered as random, the following priors were used $\mu \sim N(0, 100)$, $\sigma^{-2} \sim \Gamma(2.01/2, 0.01/2)$, $M \sim \Gamma(2.0, 0.2)$, and $c \sim \Gamma(2.0, 0.2)$, where $\Gamma(\alpha_0, \alpha_1)$ refers to a gamma distribution with scale and rate, α_0 and α_1, respectively. For the prior distribution of the difficulty parameters, a $N_{n-1}(\mathbf{0}, 10^3 \times \boldsymbol{I}_{n-1})$ prior distribution was considered. A sensitivity analysis for the choice of these hyper-parameters revealed robustness of the posterior results.

The probability distribution estimates are displays in Figure 2. The first two rows show the behavior of the estimates for the SRM, whereas the second two rows show the behavior of the corresponding estimates for the SPCRM. From left to right columns display the results for $m = 2$, $m = 4$, and $m = 40$ items, respectively. For the SPCRM, the estimates of the distribution G are quite good even for 2 probes. The situation is completely different for the SRM. When $m = 2$ items are considered, the KD and L_1 distance were approximately the double of the corresponding values for the SPCRM (see Table 1). Moreover, the posterior estimates strongly depend on the prior distribution of G. In agreement with Theorem 3, the situation

seems to improve when the number of items increases. Table 1 shows that, for the SPCRM, the eight semiparametric models outperform the normal parametric case. Furthermore, MPT priors showed the best performance.

[Figure 2 about here.]

[Table 1 about here.]

We note that these results are for one random sample from one particular density, and conclusions should be drawn with care. Nonetheless, this examples do show that although nonparametric priors are flexible and can capture different behaviors of the ability distribution, marked differences due to the identification restrictions of model parameters are observed depending on whether the data has or not an infinite support.

6.2 The SIMCE dataset

The SIMCE project in Chile has developed mandatory tests to assess regularly the educational quality in three levels: 4th and 8th grades in primary school, and 2th grade in secondary school. All students in the country are expected to take the tests which are scheduled every 3 or 4 years. For this application, data from a mathematics test (46 items with 4 alternatives) applied in 2004 to 8th graders in primary were considered. This test has a variety of binary questions ranging from problem formulation, algebra and functions, geometry, and numbers.

A significant characteristic of the chilean elementary and secondary educational system is the existence of a variety of different type of schools. They can grouped according to financing and administration in: Public I, financed by the state and administered by county governments; Public II, financed by the state and administered by county corporations; Private I, financed by the

state and administered by the private sector; Private II, fee-paying schools that operate solely on payments from parents and administered by the private sector. Table 2 summarizes general information with respect to the observed total score (the total number of correct answers) in the population.

[Table 2 about here.]

From the distribution of the examinees and the average of correct answers in Table 2, and from substantive considerations concerning the type of schools, at least a non-symmetric distributions of the abilities is expected. Thus a parametric model assuming a symmetric random effect distribution is inadequate. Although in SRM the structural parameters are not identified by a finite number of items, Theorem 3 ensures that a substantial amount of characteristics of G are identified when m is large enough. In this application, $2^{46} - 46 - 2 = 7,036.874 \times 10^{10}$ characteristics of G are identified by the data.

The original data set considers 109,981 students. For illustration purposes a random subset of 1,000 students is considered. Two DP-type and two PT-type prior specifications were considered for this data (see Table 3). For the prior distribution of the difficulty parameters, a $N_{45}(\mathbf{0}, 10^3 \times \boldsymbol{I}_{45})$ prior distribution was considered. A sensitivity analysis for the choice of these hyper-parameters revealed robustness of the posterior results. The cross validation model comparison criteria shows a clear preference toward the semiparametrics versions of the model. For instance, Table 3 shows the Pseudo-Bayes factors (PBF) (see e.g., Geisser and Eddy, 1979; Gelfand and Dey, 1994) for the model comparison of the semiparametric versions of the model versus the normal case. The PBF for model M_1 versus model M_2 is defined as $\text{PBF}_{M_1,M_2} = \prod_{i=1}^{n}\prod_{j=1}^{m} \frac{p_{M_1}(Y_{ij}|\boldsymbol{Y}^{(-ij)})}{p_{M_2}(Y_{ij}|\boldsymbol{Y}^{(-ij)})}$. The results indicated that the considered semiparametric models "predicted" approximately more than 33 times better the actual data points, in

comparison to the normal case.

[Table 3 about here.]

Figure 3 compares the posterior estimates of the cumulative distribution and density of the abilities, under the four SRM and with the results from a normal fit. Note that in the case of the DP-type priors, the density estimates are not a formal component of the model. A kernel smoothing approach, based on samples from the posterior predictive distribution of the abilities was considered in this case. The predictive abilities distributions were skewed to the right. The results supported the departure of the normality assumption of the distribution generating the individual abilities.

[Figure 3 about here.]

Finally, Table 4 presents the posterior means for five randomly sampled difficulty parameter for the SRMs, along with the results from an normal fit. The posterior means and the significance of the difficulty parameters are consistent across the models, suggesting that, for this level of disagreement between the latent variable distribution, the estimated difficulty parameters are robust against the mis-specification of the abilities distribution.

[Table 4 about here.]

6.3 French Written Test Data

We consider data from a written test constructed to perform an assessment of the curriculum standards of French as a foreign language in Flanders (Belgium). As part of a calibration study, the test was administered at the beginning of the third academic of secondary education to a

group of 483 students. The students were sampled according to a three-step design. In the first step, the schools were drawn from the population of schools using a stratified sampling scheme. In the second step, two classes were drawn at random from each school. In the final step, different test booklets were distributed among the students according to a spiraling design. The written test considered six texts, which the student had to copy. The response variable of interest was the number of spelling errors in each test. Specifically, it was coded for each word whether it was written correctly or not. Omitted words were considered as wrong. Because of the different text lengths, and based on exploratory analysis of the data, we assume that the number of spelling errors is directly proportional to the text's number of words, *i.e.*

$$(Y_{ij} \mid \lambda_{ij}) \sim \text{Poisson}(\nu_j \lambda_{ij}), \quad \lambda_{ij} = \exp(\theta_i - \beta_j) \qquad i = 1, \ldots, m, \;\; j = 1, \ldots, n, \quad (13)$$

where ν_j corresponds to the length of the text j. This kind of correction is usually referred to as an "offset" correction in the context of generalized linear models. Note that our identification result for the SPCRM is entirely valid for this version of the model. In fact, the only step we need to verify is the 2-completeness of λ_j w.r.t. Y_{ij}, which straightforward follows from the 2-completeness of the parameter of a Poisson distribution since ν_j is a fixed parameter.

The same two DP-type and two PT-type prior specifications considered for the SIMCE dataset were used for this data. Again, a sensitivity analysis for the choice of the hyper-parameters revealed robustness of the posterior results for each of the models. The cross validation model comparison criteria showed a slight preference toward the semiparametric versions of the model (see Table 5).

[Table 5 about here.]

Figure 4 displays the posterior predictive abilities distribution under the four SRMs and under the normal fit. Although the predictive cumulative distributions do not showed an important

departure from the normal model, the predictive abilities density of the semiparametric models suggested a multimodal distribution.

[Figure 4 about here.]

Table 6 presents the posterior means for the difficulty parameter for the SRPCMs, along with the results from an normal fit. The posterior means and the significance of the difficulty parameters were consistent across the models, suggesting that for the Rasch Poisson Count model, the estimated difficulty parameters are robust against the mis-specification of the abilities distribution.

[Table 6 about here.]

7 Concluding Remarks

Semiparametric IRT-type models deserve two questions: the first one deals with their empirical meaning; the second question is related with the statistical properties of the relevant estimates. In this paper, these questions were analyzed from a Bayesian point of view. The empirical meaning of the models was studied by means of a Bayesian identification analysis of the parameters of interest. When a parameter is not identified, not only part of the prior information is not revised by the observations, but also the corresponding posterior probability is always equal to the posterior distribution of some function of the corresponding identified parameter. It is important to remark that such a function is in general unknown. Although such identification issues present no difficulties to a Bayesian analysis in the sense that a prior probability is transformed into a posterior probability using the sampling model, if the interest focuses on a non identified parameter then such formal assurances have little practical value. Furthermore, when

identification is considered, structural modeling and Bayesian approach are mutually benefited in the sense that the Bayesian learning process becomes fully characterized by the parameters of interest, which in turn have a substantive meaning.

These considerations have been illustrated in both the SPCRM and SRM. The parameters of interest of the SPCRM are identified by one observation under very simple conditions; see Theorem 1. For the SRM the situation is rather different since the identification of the parameters of interest is obtained when a test is composed by an infinite number of items. When considering a test with a finite number of items, only for the SPRCM makes sense to learn about G from the data, but not for the SRM. The simulation study from Section 6.1 show this practical aspect, particularly when a test with two or four items are considered.

The statistical properties of the relevant estimates were analyzed through a genuinely Bayesian concept of consistency, rather than studying the sampling behavior of posterior expectations. We showed that the parameters of interest of the SPCRM are consistently estimated in the asymptotic experiment obtained when the number of examinees goes to infinite. Nevertheless, in agreement with the lack of identification of the SRM in a finite horizon, their parameters of interest are consistently estimated in the doubly asymptotic experiment, that is, when the number of items and the number of examinees go to infinite. Fortunately, Theorem 2 ensures coherent practical uses for the SRM since only $2^n - n - 2$ characteristics of G can be identified. When n is "large", as it is the case for the SIMCE data set analyzed in Section , the Bayesian estimates of these characteristics can carefully be used to learn "something" about the probability distribution G.

It is important to remark that our consistency results correspond to necessary conditions for the corresponding sampling consistency. Nevertheless, the more relevant aspect is that these

consistency results are straightforward consequences of the fact that the structural Rasch-type models (which is obtained after integrating out the unobserved abilities) define an *iid* process. More specifically, in an *iid* process only the Bayesian identified parameters are consistently estimated by the corresponding posterior expectations. This means that the parameters of interest are a.s. measurable functions of the infinite sequence of observations. In other words, since the parameters of interest are identified, the respective posterior expectations represent a genuinely updating process, the "limit" being to observe the parameters of interest when observing the totality of observations.

Finally, since Rasch models belong to the general class of Generalized Linear Mixed Models, we argue that identification and Bayesian consistency need to be studied in such a class of models. These extensions of our results are the subject of the current work.

Acknowledgments

The first author was partially supported by the FONDECYT Project N^{o} 1060722 from the Chilean Government. The last three authors were partially supported by the Interuniversity Attraction Poles Program P6/03 from the Belgian State Federal Office for Scientific, Technical and Cultural Affairs. The first two authors gratefully acknowledge partial support from the KUL-PUC bilateral (Flanders-Chile) grant BIL05/03. The SIMCE Office from the Chilean Government kindly allowed us access to the databases used in this paper. The authors gratefully acknowledge Rianne Jansen (KUL) for allowing access to the French written test data set.

Appendix

A Characterization of the Minimal Sufficient Parameter

Let us consider a Bayesian model $\mathcal{E} = (\Theta \times \Omega, \mathcal{A} \vee \mathcal{X}, \Pi)$, where $(\Theta, \mathcal{A})$ and $(\Omega, \mathcal{X})$ are the *parameter* and *sample* space, respectively, and Π is a unique probability measure defined on $\mathcal{A} \vee \mathcal{X}$. Let $\Sigma_{\mathcal{E}}$ be the class of sufficient parameters $\mathcal{B} \subset \mathcal{A}$ for the process generating $\mathcal{X}$, namely $\Sigma_{\mathcal{E}} = \{\mathcal{B} \subset \mathcal{A} : \mathcal{X} \perp\!\!\!\perp \mathcal{A} \mid \mathcal{B}\}$. It follows that $\mathcal{A} \in \Sigma_{\mathcal{E}}$, hence $\Sigma_{\mathcal{E}} \neq \emptyset$. Therefore, if $\mathcal{B}_1, \mathcal{B}_2 \in \Sigma_{\mathcal{E}}$, then $\overline{\mathcal{B}_1} \cap \overline{\mathcal{B}_2} \in \Sigma_{\mathcal{E}}$, where $\overline{\mathcal{B}_j}$ $(j = 1, 2)$ denotes the measurable completion $\overline{\mathcal{B}_j} = \mathcal{B}_j \vee \{E \in \mathcal{A} : \mu(E)^2 = \mu(E)\}$ and μ denotes the restriction of Π on $\mathcal{A}$ (that is, the prior distribution). Consequently, the minimal sufficient parameter $\mathcal{B}_{\text{min}} \in \Sigma_{\mathcal{E}}$ *always* exists and it is given by

$$\mathcal{B}_{\text{min}} = \bigcap_{\mathcal{B} \in \Sigma_{\mathcal{E}}} \overline{\mathcal{B}}. \tag{14}$$

Using the properties of the measurable completion (see Section 2.2.3 in Florens et al., 1990), it follows that $\overline{\mathcal{B}_{\text{min}}} = \mathcal{B}_{\text{min}}$. Thus, the minimal sufficient parameter $\mathcal{B}_{\text{min}}$ contains *all* the null sets of the parameter space $(\Theta, \mathcal{A})$ defined with respect to the prior probability μ.

The minimal sufficient parameter $\mathcal{B}_{\text{min}}$ can be expressed in more operational terms. Indeed, the σ-field generated by every version of the sampling expectations, namely $\sigma\{E(f \mid \mathcal{A}) : f \in [\mathcal{X}]^+\}$, is the smallest sub-$\sigma$-field of $\mathcal{A}$ that makes the sampling expectations measurable; here, $[\mathcal{X}]^+$ denotes the set of non-negative $\mathcal{X}$-measurable functions. This is equivalent to,

$$\mathcal{A} \perp\!\!\!\perp \mathcal{X} \mid \sigma\{E(f \mid \mathcal{A}) : f \in [\mathcal{X}]^+\}. \tag{15}$$

Therefore, the minimal sufficient parameter $\mathcal{B}_{\text{min}}$ is equal to $\sigma\{E(f \mid \mathcal{A}) : f \in [\mathcal{X}]^+\}$. We can now introduce the concept of Bayesian identification.

Definition 9 *The parameter $\mathcal{A}$ is* b-identified *by $\mathcal{X}$ if $\mathcal{A}$ is minimal sufficient (that is, $\mathcal{A} = \mathcal{B}_{min}$). More generally, let $\mathcal{M}_i \subset \mathcal{A} \vee \mathcal{X}$, for $i = 1, 2, 3$. It is said that $\mathcal{M}_1$* is b-identified by $\mathcal{M}_2$ conditionally on $\mathcal{M}_3$ *if $\mathcal{M}_1 \vee \mathcal{M}_3$ is b-identified by $\mathcal{M}_2 \vee \mathcal{M}_3$,* i.e.*,*

$$\sigma\{E(f \mid \mathcal{M}_1 \vee \mathcal{M}_3) \,:\, f \in [\mathcal{M}_2 \vee \mathcal{M}_3]^+\} \,=\, \mathcal{M}_1 \vee \mathcal{M}_3.$$

For details and properties, we refer to Chapter 4 in Florens et al. (1990).

B Proof of Theorem 1

As sketched in Section 4.1, the identification of the semiparametric Rasch Poisson model relies on three steps:

STEP 1: If $(Y \mid \lambda) \sim$ Poisson (λ), then $\lambda \ll_2 Y$ for all prior distribution $m(\lambda)$ defined on $(\mathbb{R}_+, \mathcal{B}_{\mathbb{R}_+})$.

STEP 2: For $n \geq 2$, $(\theta_i, \boldsymbol{\beta}_{2:n}) \ll_2 \boldsymbol{Y}_i \mid \beta_1$, where $\boldsymbol{Y}_i \in \mathbb{N}^n$.

STEP 3: G is identified by θ_i, it follows that

STEP 4: From Steps 2 and 3, by the Theorem 1 (p. 151) in Mouchart and San Martin (2003), it follows that $(\boldsymbol{\beta}_{2:n}, G)$ is identified by $\boldsymbol{Y}_i$ conditionally on β_1.

In what follows, we provide the details of each of these steps.

B.1 Proof of Step 1:

Let $(Y \mid \lambda) \sim$ Poisson(λ), with $Y \in \mathbb{N}$, and let $m(\lambda)$ be a probability measure on $(\mathbb{R}_+, \mathcal{B}_{\mathbb{R}_+})$. We need to prove that $\lambda \ll_2 Y$, which by definition means that, if $\forall\, h \in L^2(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+), \mu_{\lambda|Y})$,

$$\int_0^\infty h(t)\, d\mu_{\lambda|Y}(t) \,=\, 0 \implies h \,=\, 0 \;\; \mu_{\lambda|Y} - \text{a.s.} \tag{16}$$

where $\mu_{\lambda|Y}$ is the corresponding posterior probability measure on $\mathbb{R}_+$. The nullity of the integral in equation (16) is equivalent to

$$\int_0^\infty h(t)\,\frac{t^y}{y!}\,e^{-t}\,d\mu_\lambda(t) \;=\; 0, \quad \forall\, y \in \mathbb{N}. \tag{17}$$

But $L^2(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+), \mu_{\lambda|Y}) \subset L^1(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+), \mu_{\lambda|Y})$. We next use the fact that

$$\overline{\text{span}\left\{e^{-t}\frac{t^y}{y!} : y \in \mathbb{N},\ t \in \mathbb{R}_+\right\}}^{L^1(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+), \mu_{\lambda|Y})} = L^1(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+), \mu_{\lambda|Y}),$$

where the upper-bar means the closeness w.r.t. the $L^1(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+), \mu_{\lambda|Y})$-norm; for a proof of this last equality, see San Martín (2007). By a duality argument, it follows that expression (17) implies that $h \;=\; 0 \;\; \mu_\lambda -$ a.s.

■

B.2 Proof of Step 2:

The underlying arguments are based on the following theorem established in Florens et al. (1990).

Theorem 5 *Let $p \in [1, \infty]$ and let X_1, X_2, X_3 and X_4 be random variables defined on a common probability space $(\Omega, \mathcal{M}, P)$. If $X_2 \perp\!\!\!\perp X_4 \mid (X_1, X_3)$ then*

(i) $X_1 \ll_p X_2 \mid X_3$ *implies* $X_1 \ll_p X_2 \mid (X_3, X_4)$ *(Florens et al. 1990, Theorem 5.4.5).*

(ii) If $X_4 \ll_p X_1 \mid X_3$ *and* $X_1 \ll_p X_2 \mid X_3$*, then* $X_4 \ll_p X_2 \mid X_3$ *(Florens et al. 1990, Theorem 5.4.10).*

The following lemma is used in the next steps of the proof.

Lemma 1 *Let X_1, X_2, X_3, X_4, X_5, be random variables defined on a common probability space $(\Omega, \mathcal{M}, P)$. The following conditions (with $p \in [1, \infty]$)*

$$\textit{(i)} \quad X_1 \ll_p X_2 \mid X_5, \qquad \textit{(ii)} \quad X_3 \ll_p X_4 \mid X_5$$

jointly imply that $(X_1, X_3) \ll_p (X_2, X_4) \mid X_5$ under the following two conditions:

$$\textit{(iii)} \quad X_2 \perp\!\!\!\perp X_3 \mid (X_1, X_5), \qquad \textit{(iv)} \quad X_4 \perp\!\!\!\perp (X_1, X_2) \mid (X_3, X_5).$$

Proof of Lemma B.1: Let $p \in [1, \infty]$. Since $X_1 \ll_p X_2 \mid X_5$ is equivalent to $(X_1, X_5) \ll_p (X_3, X_5)$, it is sufficient to prove this lemma for $X_5 = E(X_5)$ a.s. By Theorem 5.i, $X_1 \ll_p X_2$ along with $X_2 \perp\!\!\!\perp X_3 \mid X_1$ imply (v) $(X_1, X_3) \ll_p (X_2, X_3)$. Similarly, $X_3 \ll_p X_4$ along with $X_2 \perp\!\!\!\perp X_4 \mid X_3$ (a property implied by condition (iv) above) imply (vi) $(X_2, X_3) \ll_p (X_2, X_4)$. Since $X_1 \perp\!\!\!\perp X_4 \mid (X_2, X_3)$ (a property implied by condition (iv) above), conditions (v) and (vi) jointly imply that $(X_1, X_3) \ll_p (X_2, X_4)$ by Theorem 5.ii.

■

Remark 3 *Conditions (iii) and (iv) of Lemma 1 are implied by the stronger condition $(X_1, X_2) \perp\!\!\!\perp (X_3, X_4) \mid X_5$. Moreover, under this condition, properties (i) and (ii) of Lemma 1 are jointly equivalent to $(X_1, X_3) \ll_p (X_2, X_4) \mid X_5$.*

Assumption (4) implies that $Y_{ij} \perp\!\!\!\perp (\theta_i, \beta_j) \mid \theta_i - \beta_j$ for all $i = 1, \ldots, m$ and for all $j = 1, \ldots, n$. Since $\theta_i - \beta_j$ is a function of (θ_i, β_j), the latter condition, along with H2.iv, imply that

$$Y_{ij} \perp\!\!\!\perp (\theta_i, \boldsymbol{\beta}_{1:n}) \mid \theta_i - \beta_j \qquad \forall\, i = 1, \ldots, m, \;\; j = 1, \ldots, n. \tag{18}$$

By Step 1, $\theta_i - \beta_j \ll_2 Y_{ij}$ for all $j = 1, \ldots, n$. This condition, along with $Y_{ij} \perp\!\!\!\perp \beta_1 \mid \theta_i - \beta_j$ (a property implied by (18)), imply by Theorem 5.i, that $\theta_i - \beta_j \ll_2 Y_{ij} \mid \beta_1$ for all $j = 2, \ldots, n$.

Let us now suppose that, for $l < n$,

$$(\theta_i - \beta_1, \ldots, \theta_i - \beta_l) \ll_2 (Y_{i1}, \ldots, Y_{il}) \mid \beta_1.$$

Using Lemma 1, this condition, along with $\theta_i - \beta_{l+1} \ll_2 Y_{i,l+1} \mid \beta_1$, imply that

$$(\theta_i - \beta_1, \ldots, \theta_i - \beta_l, \theta_i - \beta_{l+1}) \ll_2 (Y_{i1}, \ldots, Y_{il}, Y_{i,l+1}) \mid \beta_1,$$

provided that,

$$(Y_{i1}, \ldots, Y_{il}) \perp\!\!\!\perp (\theta_i - \beta_{l+1}) \mid (\theta_i - \beta_1, \ldots, \theta_i - \beta_l, \beta_1),$$

$$Y_{i,l+1} \perp\!\!\!\perp (\theta_i - \beta_1, \ldots, \theta_i - \beta_l, Y_{i1}, \ldots, Y_{il}) \mid (\theta_i - \beta_{l+1}, \beta_1).$$

It is straightforward to verify that these last conditions are implied by properties H2.iii, H2.iv and H2.v. Consequently, by induction on l, it follows that

$$(\theta_i - \beta_1, \ldots, \theta_i - \beta_n) \ll_2 (Y_{i1}, \ldots, Y_{in}) \mid \beta_1. \tag{19}$$

But

$$\begin{pmatrix} \mathbf{1}_n & -I_n \\ 0 & e_1' \end{pmatrix} \begin{pmatrix} \theta_i \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} = \begin{pmatrix} \theta_i - \beta_1 \\ \vdots \\ \theta_i - \beta_n \\ \beta_1 \end{pmatrix}, \tag{20}$$

where $\mathbf{1}_n = (1, \ldots, 1)' \in \mathbb{R}^n$ and e_1 is the first canonical vector of $\mathbb{R}^n$. Since the matrix in (20) is non-singular, condition (19) is equivalent to,

$$(\theta_i, \boldsymbol{\beta}_{2:n}) \ll_2 \boldsymbol{Y}_i \mid \beta_1. \tag{21}$$

■

B.3 Proof of Step 3:

Since $(\theta_i \mid G) \sim G$, G is the minimal sufficient parameter for θ_i. Since $\theta_i \perp\!\!\!\perp \beta_1 \mid G$, a property implied by H3.i and H3.ii, it follows that G is b-identified by θ_i conditionally on β_1.

■

B.4 Proof of Step 4:

The proof is completed by using the Theorem 1 (p. 151) in Mouchart and San Martin (2003).

Theorem 6 (Mouchart and San Martin, 2003, Theorem 1) *Consider the following statistical model,*

$$p(\boldsymbol{Y}_{1:m} \mid \boldsymbol{\eta}_1, \boldsymbol{\eta}_2) = \int p(\boldsymbol{Y}_{1:m} \mid \boldsymbol{\eta}_1, \boldsymbol{\theta}_{1:m}) p(\boldsymbol{\theta}_{1:m} \mid \boldsymbol{\eta}_2)\, d\boldsymbol{\theta}_{1:m},$$

where $\boldsymbol{Y}_{1:m} \perp\!\!\!\perp \boldsymbol{\eta}_2 \mid \boldsymbol{\eta}_1, \boldsymbol{\theta}_{1:m}$ *and* $\boldsymbol{\theta}_{1:m} \perp\!\!\!\perp \boldsymbol{\eta}_1 \mid \boldsymbol{\eta}_2$*. If* $\boldsymbol{\eta}_2$ *is* b*-identified by* $\boldsymbol{\theta}_{1:m}$ *and if* $(\boldsymbol{\eta}_1, \boldsymbol{\theta}_{1:m})$ *is 2-strongly identified by* $\boldsymbol{Y}_{1:m}$*, then* $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ *is* b*-identified by* $\boldsymbol{Y}_{1:m}$*.*

From steps 1-3, it follows that $(\boldsymbol{\beta}_{2:n}, G)$ are b-identified by $\boldsymbol{Y}_i$ conditionally on β_1.

■

C Proof of Theorem 3

Let $\mathcal{A}^* = \sigma\{E[f \mid \boldsymbol{\beta}_{1:\infty}, G] \ : \ f \in \sigma(\boldsymbol{Y}_1)\}$ be the minimal sufficient parameter of the asymptotic Bayesian Rasch model obtained as the limit when $n \to \infty$. As sketched in section 4.2, the identification of the SRM relies on the following steps:

STEP 1: For each $n \in \mathbb{N}$ and $j = 2, \ldots, n$, $\sigma(\boldsymbol{\delta}_{2:\infty}) \doteq \sigma\{\delta_j \ : \ j \geq 2\} \subset \mathcal{A}^*$.

STEP 2: Let $H_{\beta_1}((-\infty, x]) \doteq H((-\infty, x + \beta_1])$ for $x \in \mathbb{R}$.

2.A: H_{β_1} is $\overline{\mathcal{A}^*}$-measurable, where the upper-bar denotes a σ-field completed with measurable null sets; for a definition, see Section A (Appendix).

2.B: $E[g\{K(\theta_1, \beta_1)\} \mid \boldsymbol{\beta}_{1:\infty}, G]$ is $\overline{\mathcal{A}^*}$-measurable, where $K(\theta_1, \beta_1) \doteq \int_{\mathbb{R}} \Psi(\theta_1 - \beta_1 - \delta)\, H_{\beta_1}(d\delta)$.

2.C: Let $G_{\beta_1}((-\infty, x]) \doteq G((-\infty, x + \beta_1])$ for $x \in \mathbb{R}$. It follows that $\sigma(G_{\beta_1}) \subset \overline{\mathcal{A}^*}$.

STEP 3: It can be concluded that $\sigma(G_{\beta_1}) \vee \sigma(\boldsymbol{\delta}_{2:\infty}) \subset \overline{\mathcal{A}^*}$. Adding a restriction ensuring that $\sigma(G_{\beta_1}) = \sigma(G) \vee \sigma(\beta_1)$, it follows that $\sigma(G) \vee \sigma(\boldsymbol{\beta}_{1:\infty}) \subset \overline{\mathcal{A}^*}$.

Before providing the details of each of these steps, let us introduce the following assumption:

$$\text{(i)} \underset{1 \le j < \infty}{\perp\!\!\!\perp} \beta_j \mid H, \qquad \text{(ii)} \ (\beta_j \mid H) \sim H \quad \forall\, j \ge 1. \tag{22}$$

It is also assumed that H satisfies properties H3 and H4 where $H_\infty \equiv H$.

C.1 Proof of Step 1

By the same argument used to establish equality (11), hypothesis H2.iii implies $\forall\, J \subset n\{j \ : \ 1 \le j \le n\}$ that

$$P\left[\bigcap_{j \in J} \{Y_{ij} = 1\} \cap \bigcap_{j \in J^c} \{Y_{ij} = 0\} \mid \boldsymbol{\beta}_{1:\infty}, G\right] = e^{-\sum_{j \in J} \beta_j} \times \int_{\mathbb{R}} \frac{e^{|J|\theta}}{\prod_{1 \le j \le m} \{1 + e^{\theta - \beta_j}\}}\, G(d\theta)$$

where $|J|$ is the cardinality of set J. By taking the above relation with $J = \{j\}$ divided by the same relation with $J = \{1\}$, it follows that, for all $2 \le j \le n$, $\delta_j \doteq \beta_j - \beta_1$ is $\mathcal{A}^*$-measurable. Therefore, $\sigma(\boldsymbol{\delta}_{2:\infty}) \doteq \sigma\{\delta_j \ : \ j \ge 2\} \subset \mathcal{A}^*$.

■

C.2 Proof of Step 2.A

Condition (22.i) clearly implies that $\perp\!\!\!\perp_{2\le j<\infty}(\beta_j, \beta_1) \mid (H, \beta_1)$. Since δ_j is a function of (β_1, β_j), it follows that $\perp\!\!\!\perp_{2\le j<\infty} \delta_j \mid (H, \beta_1)$. Moreover, (22) implies that $\boldsymbol{\beta}_{1:\infty} \perp\!\!\!\perp G \mid H$, which in turn implies that $\boldsymbol{\delta}_{2:\infty} \perp\!\!\!\perp G \mid (H, \beta_1)$. Therefore, $(\delta_j \mid H, \beta_1) \sim (\delta_2 \mid H, \beta_1)$ for all $j > 2$. Finally, for all $x \in \mathbb{R}$,

$$\begin{aligned} P\{\delta_j \in (-\infty, x] \mid H, \beta_1\} &= P\{\beta_j \in (-\infty, x+\beta_1] \mid H, \beta_1\} \\ &= H((-\infty, x+\beta_1]) \qquad (23) \\ &\doteq H_{\beta_1}((-\infty, x]). \qquad (24) \end{aligned}$$

Thus, $\{\delta_j \ : \ j \ge 2\}$ forms an *iid* process conditionally on H_{β_1}. Using Theorem 4, it follows that the corresponding minimal sufficient parameter, namely $\sigma\{E(d \mid H, \beta_1) \ : \ d \in [\sigma(\delta_2)]^+\}$, is a.s. a function of the infinite sequence $(\delta_2, \delta_3 \dots)$. But $(\delta_2 \mid H, \beta_1) \sim H_{\beta_1}$, which implies that H_{β_1} is b-identified by δ_2. Therefore, H_{β_1} is $\overline{\sigma(\boldsymbol{\delta}_{2:\infty})}$-measurable. Using Step 1, the Step 2 is proved.

■

C.3 Proof of Step 2.B

Let $S_{1n} = \sum_{1\le j\le n} Y_{1j}$. Then

$$E\left(\tfrac{S_{1n}}{n} \mid \boldsymbol{\beta}_{1:\infty}, G, \theta_1\right) = \tfrac{1}{n} \sum_{1\le j\le n} \Psi(\theta_1 - \beta_1 - \delta_j),$$

$$V\left(\tfrac{S_{1n}}{n} \mid \boldsymbol{\beta}_{1:\infty}, G, \theta_1\right) = \tfrac{1}{n^2} \sum_{1\le j\le n} \Psi(\theta_1 - \beta_1 - \delta_j)\left\{1 - \Psi(\theta_1 - \beta_1 - \delta_j)\right\}.$$

Therefore, conditionally on (θ_1, β_1, H), by the Strong Law of Large Numbers it follows that

$$P\left\{\lim_{n\to\infty} E\left(\frac{S_{1n}}{n} \mid \boldsymbol{\beta}_{1:\infty}, G, \theta_1\right) = K(\theta_1, \beta_1) \,\middle|\, \theta_1, \beta_1, H\right\} = 0,$$

where $K(\theta_1, \beta_1) \doteq \int_{\mathbb{R}} \Psi(\theta_1 - \beta_1 - \delta) H_{\beta_1}(d\delta)$. By taking expectation w.r.t. (θ_1, β_1, H), this convergence is also marginally almost sure, and therefore, as $n \to \infty$,

$$E\left(\left\{\frac{S_{1n}}{n} - K(\theta_1, \beta_1)\right\}^2 \mid \boldsymbol{\beta}_{1:\infty}, G, \theta_1\right) \longrightarrow 0 \quad \text{a.s.}$$

Now, for every g bounded continuous function on $(0, 1)$, we also have

$$E\left[\left\{g\left(\frac{S_{1n}}{n}\right) - g[K(\theta_1, \beta_1)]\right\}^2 \mid \boldsymbol{\beta}_{1:\infty}, G, \theta_1\right] \longrightarrow 0 \quad \text{a.s.} \quad \text{as } n \to \infty.$$

By the Dominated Conditional Convergence Theorem, we obtain that

$$E\left[\left\{g\left(\frac{S_{1n}}{n}\right) - g[K(\theta_1, \beta_1)]\right\}^2 \mid G, \beta_1^{\infty}\right] \longrightarrow 0 \quad \text{a.s.} \quad \text{a.s.} \quad \text{as } n \to \infty.$$

This implies that, as $n \to \infty$,

$$E\left[g\left(\frac{S_{1n}}{n}\right) \mid \boldsymbol{\beta}_{1:\infty}, G\right] \longrightarrow E[g\{K(\theta_1, \beta_1) \mid \boldsymbol{\beta}_{1:\infty}, G\}] \quad \text{a.s.}$$

Therefore, $E[g\{K(\theta_1, \beta_1) \mid \boldsymbol{\beta}_{1:\infty}, G\}]$ is $\overline{\mathcal{A}^*}$-measurable for every g bounded continuous function on $(0, 1)$.

■

C.4 Proof of Step 2.C

For $x \in \mathbb{R}$, let $G_{\beta_1}((-\infty, x]) \doteq G((-\infty, x + \beta_1])$, which is well defined since, by H3.i and H3.ii, $\theta_1 \perp\!\!\!\perp \beta_1 \mid G$. Then

$$\begin{aligned} E[g\{K(\theta_1, \beta_1) \mid \boldsymbol{\beta}_{1:\infty}, G\}] &= \int_{\mathbb{R}} g[K(\theta_1, \beta_1)]\, G(d\theta) \\ &= \int_{\mathbb{R}} g\left[\int_{\mathbb{R}} \Psi(\theta - \beta_1 - \delta) H_{\beta_1}(d\delta)\right] G(d\theta) \\ &= \int_{\mathbb{R}} g\left[L(x)\right] G_{\beta_1}(dx), \end{aligned}$$

where $L(x) \doteq \int_{\mathbb{R}} \Psi(x-\delta) H_{\beta_1}(d\delta)$ is a strictly increasing function that is known because it is measurable with respect to $\sigma(H_{\beta_1})$.

Using Step 2.A, it follows that $\sigma(H_{\beta_1}) \subset \overline{\sigma(\boldsymbol{\delta}_{2:\infty})} \subset \overline{\mathcal{A}^*}$. Hence, for every f bounded continuous function on $\mathbb{R}$, by taking $g(y) = f\left[\bar{L}(y)\right]$ where $\bar{L}(y) \doteq \inf\{x \,:\, L(x) > y\}$, we obtain that $\int_{\mathbb{R}} f(x) G_{\beta_1}(dx)$ is $\overline{\mathcal{A}^*}$-measurable. Considering

$$f_n(y) = \mathbb{1}_{(-\infty,x]}(y) + [1 - n(y-x)]\mathbb{1}_{(x,x+\frac{1}{n})}(y) \downarrow \mathbb{1}_{(-\infty,x]}(y) \qquad \forall\, y \in \mathbb{R},$$

as $n \to \infty$, the Monotone Convergence Theorem implies that, for all $x \in \mathbb{R}$, $G_{\beta_1}((-\infty, x])$ is $\overline{\mathcal{A}^*}$-measurable. Therefore, $\sigma(G_{\beta_1}) \subset \overline{\mathcal{A}^*}$.

■

C.5 Proof of Step 3

From Steps 1 and 2.C, it follows that $\sigma(G_{\beta_1}) \vee \sigma(\boldsymbol{\delta}_{2:\infty}) \subset \overline{\mathcal{A}^*}$. Finally, it is sufficient to find a condition to ensure that $\sigma(G_{\beta_1}) = \sigma(G) \vee \sigma(\beta_1)$. For instance, if G have a.s. a fix location, either a p-quantile, *i.e.* x_p for $p \in (0,1)$ such that

$$\overline{G}(p) = \inf\{x \,:\, G(x) > p\} = \inf\{x \,:\, G(x) \geq p\} = x_p,$$

or mean μ. Taking into account that $\{\beta_1, \delta_j \,:\, j \geq 2\}$ is in a one to one correspondence with $\boldsymbol{\beta}_{1:\infty}$, we get

$$\sigma(G_{\beta_1}) \vee \sigma(\boldsymbol{\delta}_{2:\infty}) = \sigma(G) \vee \sigma(\beta_1) \vee \sigma(\boldsymbol{\delta}_{2:\infty}) = \sigma(G) \vee \sigma(\boldsymbol{\beta}_{1:\infty}) \subset \overline{\mathcal{A}^*}.$$

Alternatively, if β_1 is known, that is, conditionally on β_1, it follows that $(\boldsymbol{\beta}_{1:\infty}, G)$ is b-identified by $\boldsymbol{Y}_1$.

■

References

Agresti, A., B. Caffo, and P. Ohman-Strickland (2004). Examples in which misspecification of a random effects distribution reduces efficiency. *Computational Statistics and Data Analysis 47*, 639–653.

Andersen, E. B. (1980). *Discrete Statistical Models with Social Sciences Applications*. North-Holland.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics 2*, 1152–1174.

Barankin, E. W. (1961). Sufficient parameters: solution to the minimal dimensionality problem. *Annals of the Institute of Statistical Mathematics 12*, 91–118.

Barankin, E. W., H. Kud¯ o, and T. Kusama (1980). Specification of Statistical Models by Sufficiency. In K. Matusita (Ed.), *Developments in Statistical Inference and Data Analysis*, pp. 9–19. North Holland.

Bartholomew, D. J. (1987). *Latent Variable Models and Factor Analysis*. Charles Griffin.

Bechger, T. M., N. D. Verhelst, and H. H. F. M. Verstralen (2001). Identifiability of nonlinear logistic test models. *Psychometrika 66*, 357–372.

Borsboom, D., G. J. Mellenbergh, and J. van Heerden (2003). The Theoretical Status of Latent Variables. *Psychological Review 110*, 203–219.

Bush, C. A. and S. N. MacEachern (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika 83*, 275–285.

Chandra, S. (1977). On the Mixture of Probability Distributions. *Scandinavian Journal of Statistics 4*, 105–112.

Conway, J. B. (1985). *A Course in Functional Analysis*. Springer.

Dawid, A. P. (1979). Conditional independence in statistical theory (with discussion). *Journal of the Royal Statistical Society, Series B 41*, 1–31.

De Boeck, P. and M. Wilson (2004). *Explanatory Item Response Models. A Generalized Linear and Nonlinear Approach*. Springer.

De Leeuw, J. and N. Verhelst (1986). Maximum Likelihood Estimation in Generalized Rasch Models. *Journal of Educational Statistics 11*, 183–196.

Diaconis, P. and D. Freedman (1986). On the Consistency of Bayes Estimates (with Discussion). *The Annals of Statistics 14*, 1–26.

Escobar, M. D. and M. West (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association 90*, 577–588.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics 1*, 209–230.

Ferguson, T. S. (1974). Prior distribution on the spaces of probability measures. *The Annals of Statistics 2*, 615–629.

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of Normal distributions. In D. Siegmund, J. Rustage, and G. G. Rizvi (Eds.), *Recent Advances in Statistics: Papers in Honor of Herman Chernoff on His Sixtieth Birthday*, pp. 287–302. Bibliohound.

Fischer, R. A. (1922). On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society of London, Series A 222*, 309–368.

Florens, J.-P. and M. Mouchart (1986). Exhaustivité, ancillarité et identification en statistique bayesienne. *Annales d'Économetrie et de Statistique 4*, 63–93.

Florens, J.-P., M. Mouchart, and J.-M. Rolin (1990). *Elements of Bayesian Statistics*. Marcel Dekker.

Florens, J.-P. and J.-M. Rolin (1984). Asymptotic sufficiency and exact estimability. In J.-P. Florens, M. Mouchart, J.-P. Raoult, and L. Simar (Eds.), *Alternative Approaches to Time Series Analysis*, pp. 121–142. Publications des Facultés Universitaires Saint-Louis, Bruxelles.

Geisser, S. and W. Eddy (1979). A predictive approach to model selection. *Journal of the American Statistical Association 74*, 153–160.

Gelfand, A. E. and D. Dey (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B 56*, 501–514.

Ghosh, M. (1995). Inconsistent maximum likelihood estimators for the Rasch model. *Statistical and Probability Letters 23*, 165–170.

Hanson, T. (2006). Inference for Mixtures of Finite Polya Tree Models. *Journal of the American Statistical Association 101*, 1548–1565.

Hanson, T. and W. O. Johnson (2002). Modeling regression error with a mixture of Polya trees. *Journal of the American Statistical Association 97*, 1020–1033.

Heagerty, P. J. and B. F. Kurland (2001). Misspecified maximum likelihood estimates and generalized linear mixed models. *Biometrika 88*, 973–985.

Jansen, M. G. H. (1994). Parameters of the Latent Distribution in Rasch's Poisson Counts Model. In G. H. Fischer and D. Laming (Eds.), *Contribution to Mathematical Psychology, Psychometrics, and Methodology*, pp. 319–326. Springer.

Jansen, M. G. H. and M. A. J. van Dujin (1992). Extensions of Rasch's Multiplicative Poisson Model. *Psychometrika 57*, 405–414.

Jara, A. (2007). Applied Bayesian Non- and Semi-parametric Inference using DPpackage. *Rnews 7*(3), 17–26.

Jara, A., T. Hanson, and E. Lesaffre (2007). Robustifying Generalized Linear Mixed Models using Mixtures of Multivariate Polya Trees. *Submitted*.

Kadane, J. (1974). The role of identification in Bayesian theory. In S. Fienberg and A. Zellner (Eds.), *Studies in Bayesian Econometrics and Statistics*, pp. 175–191. North Holland.

Kiefer, J. and J. Wolfowitz (1956). Consistency of the Maximum Likelihood Estimators in the Presence of Infinitely Many Incidental Parameters. *The Annals of Mathematical Statistics 27*, 887–906.

Kleinman, K. P. and J. G. Ibrahim (1998a). A Semi-Parametric Bayesian Approach to Generalized Linear Mixed Models. *Statistics in Medicine 17*, 2579–2596.

Kleinman, K. P. and J. G. Ibrahim (1998b). A Semiparametric Bayesian Approach to the Random Effects Models. *Biometrics 54*, 921–938.

Koopmans, T. C. and O. Reiersøl (1950). The Identification of Structural Characteristics. *The Annals of Mathematical Statistics 21*, 165–181.

Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modeling. *The Annals of Statistics 20*, 1222–1235.

Lavine, M. (1994). More aspects of Polya tree distributions for statistical modeling. *The Annals of Statistics 22*, 1161–1176.

Lindley, D. V. (1971). *Bayesin Statistics: A Review*. Society for Industrial and Applied Mathematics.

Lindsay, B. G., C. C. Clogg, and J. Grego (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model in item analysis. *Journal of the American Statistical Association 86*, 96–107.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates I: Density estimates. *The Annals of Statistics 12*, 351–357.

Lord, F. M. (1969). Estimating True-Score Distributions in Psychologycal Testing (An Empirical Bayes Estimation Problem). *Psychometrika 34*, 259–299.

Mauldin, R. D., W. D. Sudderth, and S. C. Williams (1992). Polya trees and random distributions. *The Annals of Statistics 20*, 1203–1221.

Mouchart, M. and J. M. Rolin (1984). A note on conditional independence with statistical applications. *Statistica 44*, 557–584.

Mouchart, M. and E. San Martin (2003). Specification and identification issues in models involving a latent hierarchical structure. *Journal of Statistical Planning and Inference 111*, 143–163.

Mukhopadhyay, S. and A. E. Gelfand (1997). Dirichlet process mixed generalized linear models. *Journal of the American Statistical Association 92*, 633–647.

Neyman, J. and E. Scott (1948). Consistent Estimates Based on Partially Consistent Observations. *Econometrica 16*, 1–32.

Pfanzagl, J. (1970). Consistent Estimation in the Presence of Incidental Parameters. *Metrika 15*, 141–148.

Pfanzagl, J. (1993). Incidental versus Random Nuisance Parameters. *The Annals of Statistics 21*, 1663–1691.

Picci, G. (1977). Some connections between the theory of sufficient statistics and the identifiability problem. *SIAM Journal on Applied Mathematics 33*, 383–398.

R Development Core Team (2006). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. The Danish Institute for Educational Research (Expanded Edition, 1980, Chicago: The University Chicago Press).

Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica 39*, 577–591.

San Martín, E. (2007). A Proof of the L_1-dense Property of Gamma Polynomials. Technical report, Pontifica Universidad Católica de Chile, Department of Statistics.

San Martín, E. and M. Mouchart (2007). On Joint Completness: Sampling and Bayesian versions, and their Connections. *Sankhyā. To appear*.

Teicher, H. (1961). Identifiability of mixtures. *The Annals of Statistics 32*, 244–248.

Verbeke, G. and E. Lesaffre (1996). A linear mixed-effects model with heterogeneity in the random effects population. *Journal of the American Statistical Association 91*, 217–221.

Walker, S. G. and B. K. Mallick (1997). Hierarchical Generalized Linear Models and Frailty Models with Bayesian Nonparametric Mixing. *Journal of the Royal Statistical Society, Ser. B 59*, 845–860.

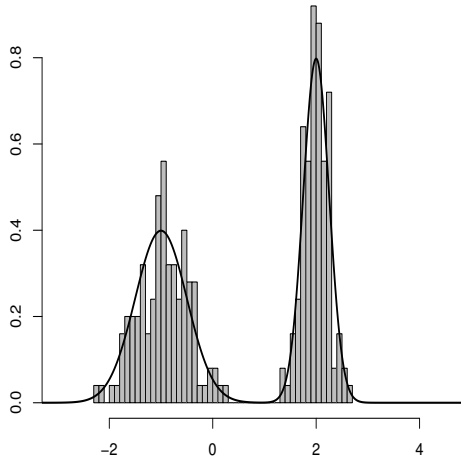


Figure 1: Simulated data set: True density and histogram of $n = 250$ simulated abilities.

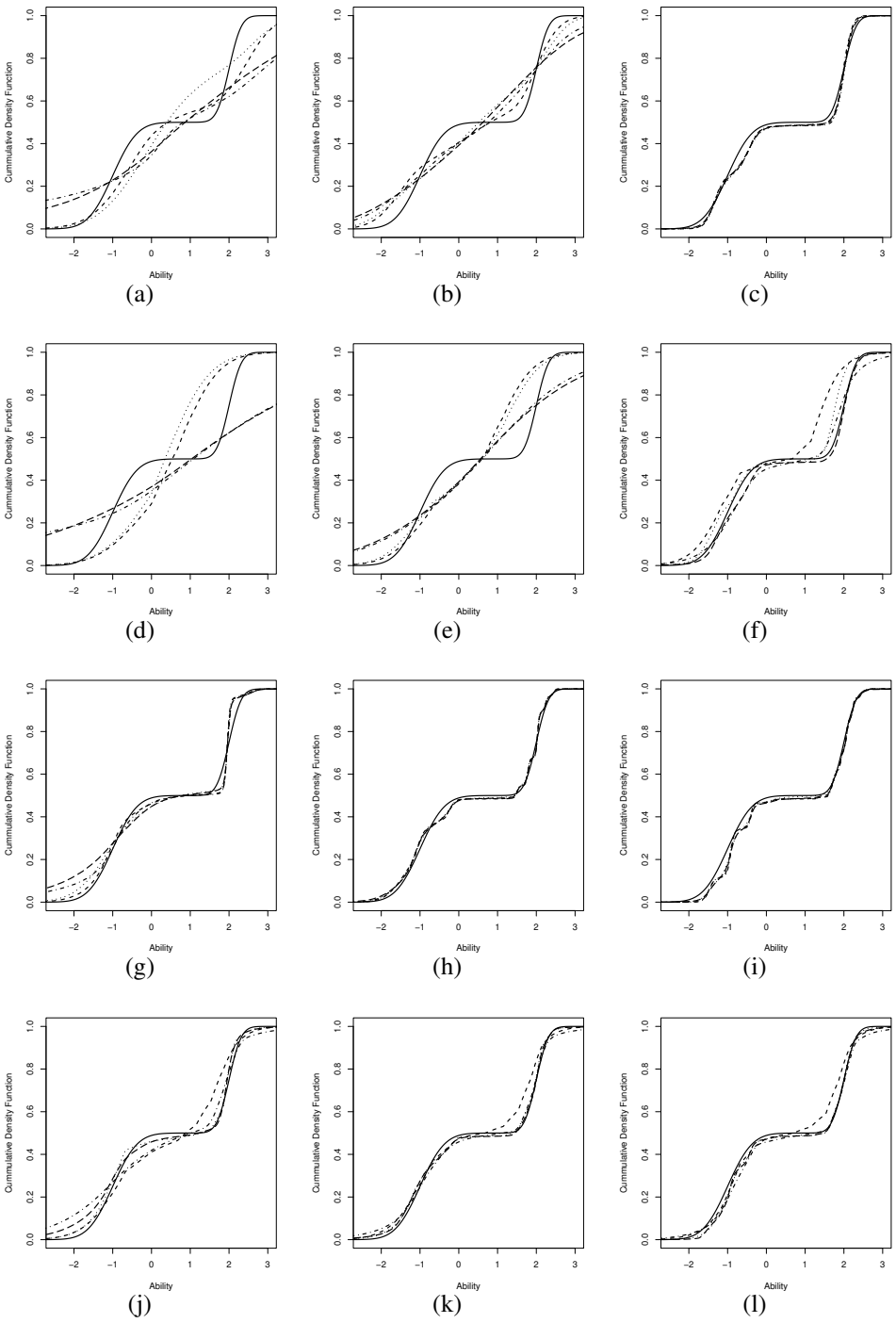


Figure 2: Simulated Datasets: Posterior predictive distribution for different semiparametric Rasch (panels (a)-(f)) and Poisson count Rasch models (panels (g)-(l)). The first-third and second-fourth rows display the results for the DP- and PT-type of priors. From left to right panels display the results for 2, 4, and 40 items/probes, respectively. In all cases, the solid line represents the true distribution. The dashed, dotted, doted-dashed and long-dashed represent the DP$(1, N(0,1))$ or PT$(\Pi^{0,1}, \mathcal{A}^1)$, MDP$(M, N(0,1))$ or MPT$(\Pi^{0,1}, \mathcal{A}^c)$, MDP$(1, N(\mu, \sigma^2))$ or MPT$\left(\Pi^{\mu,\sigma^2}, \mathcal{A}^1\right)$, and MDP$(M, N(\mu, \sigma^2))$ or MPT$\left(\Pi^{\mu,\sigma^2}, \mathcal{A}^c\right)$ version of the model, respectively.

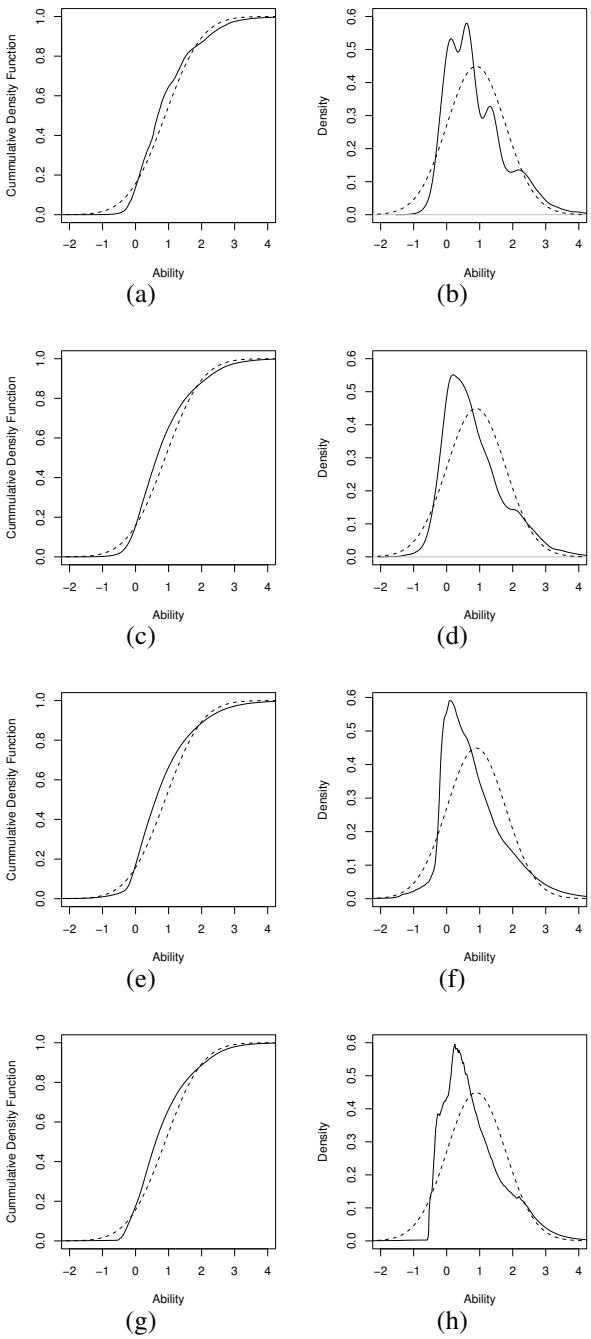


Figure 3: SIMCE Dataset: Posterior predictive abilities distribution and density, resulting from a $\int DP\left(N(\mu,\sigma^2)\right)P(d\mu,d\sigma^2)$ (panels (a) and (b)), $\int DP\left(MN(\mu,\sigma^2)\right)P(dM,d\mu,d\sigma^2)$ (panels (c) and (d)), $\int PT(\Pi^{\mu,\sigma^2},A^1)P(d\mu,d\sigma^2)$ (panels (e) and (f)), and $\int PT(\Pi^{\mu,\sigma^2},A^c)P(dc,d\mu,d\sigma^2)$ (panels (g) and (h)) prior. The nonparametric posterior predictive distributions are presented in solid lines. The parametric normal estimation is presented in the short-dashed line for comparison purposes.

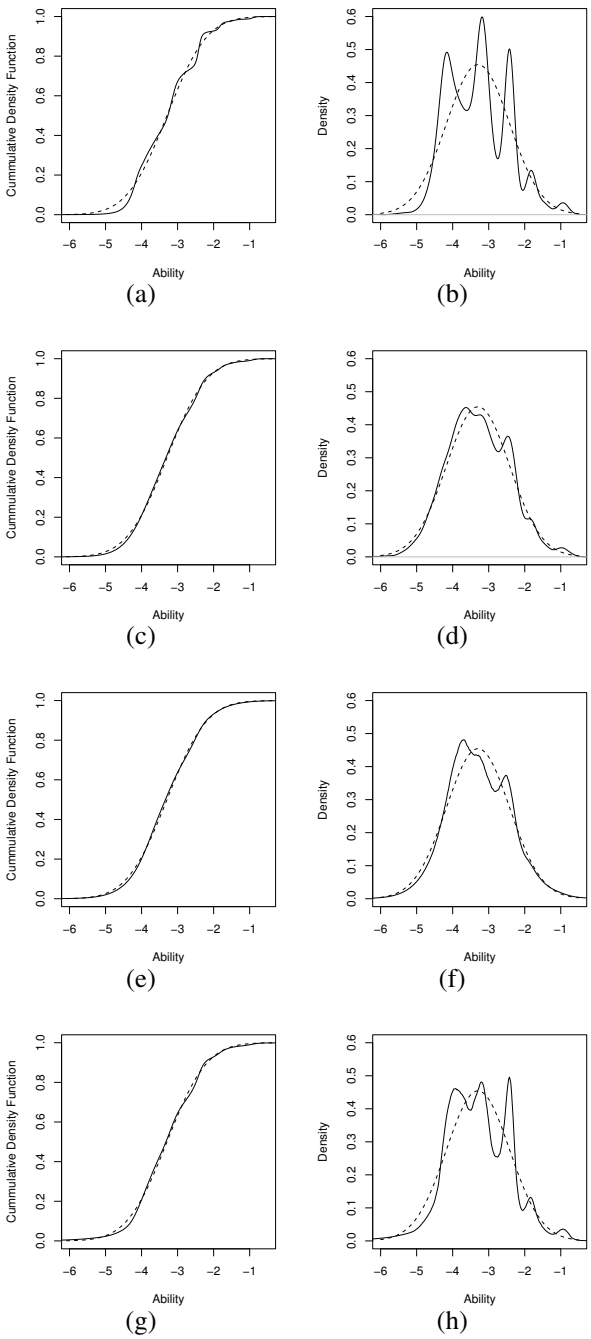


Figure 4: French Written Test Dataset: Posterior predictive abilities distribution and density, resulting from a $\int DP\left(N(\mu,\sigma^2)\right)P(d\mu,d\sigma^2)$ (panels (a) and (b)), $\int DP\left(MN(\mu,\sigma^2)\right)P(dM,d\mu,d\sigma^2)$ (panels (c) and (d)), $\int PT(\Pi^{\mu,\sigma^2},A^c)P(d\mu,d\sigma^2)$ (panels (e) and (f)), and $\int PT(\Pi^{\mu,\sigma^2},A^c)P(dc,d\mu,d\sigma^2)$ (panels (g) and (h)) prior. The nonparametric posterior predictive distributions are presented in solid lines. The parametric normal estimation is presented in the short-dashed line for comparison purposes.

Table 1: Simulated Datasets: Comparison of L_1 and Kolmogorov distance (KD), and log-Pseudo Marginal Likelihood (LPML) for different semiparametric Rasch and Poisson count Rasch models, under three different number n of items/probes.

		Rasch Model			**Rasch Poisson Model**		
Prior	n	L_1	KD	LPML	L_1	KD	LPML
$N(\mu, \sigma^2)$	2	0.100	0.305	-253.13	0.090	0.222	-692.75
	4	0.056	0.176	-447.75	0.087	0.227	-1902.64
	40	0.131	0.208	-4668.54	0.053	0.218	-15768.31
$DP(M = 1, N(0, 1))$	2	-	0.213	-254.42	-	0.130	-632.92
	4	-	0.099	-451.57	-	0.061	-1882.02
	40	-	0.077	-4598.84	-	0.092	-15779.34
$MDP(M, N(0, 1))$	2	-	0.194	-256.30	-	0.120	-631.93
	4	-	0.149	-450.79	-	0.060	-1882.75
	40	-	0.076	-4599.94	-	0.067	-15773.97
$MDP(M = 1, N(\mu, \sigma^2)$	2	-	0.300	-258.33	-	0.142	-633.50
	4	-	0.139	-451.22	-	0.056	-1886.89
	40	-	0.079	-4599.62	-	0.103	-15775.09
$MDP(M, N(\mu, \sigma^2)$	2	-	0.260	-251.50	-	0.124	-636.87
	4	-	0.158	-450.16	-	0.051	-1879.94
	40	-	0.070	-4601.25	-	0.077	-15769.48
$PT(\Pi^{0,1}, \mathcal{A}^1)$	2	0.101	0.355	-264.61	0.052	0.178	-653.88
	4	0.049	0.329	-457.88	0.036	0.131	-1876.93
	40	0.102	0.294	-4619.35	0.020	0.118	-15763.77
$MPT(\Pi^{0,1}, \mathcal{A}^c)$	2	0.102	0.391	-265.97	0.031	0.062	-624.70
	4	0.046	0.291	-457.83	0.026	0.041	-1873.00
	40	0.066	0.172	-4603.15	0.013	0.053	-15760.62
$MPT(\Pi^{\mu,\sigma^2}, \mathcal{A}^1)$	2	0.099	0.306	-252.19	0.049	0.113	-628.56
	4	0.055	0.169	-449.57	0.024	0.042	-1872.41
	40	0.053	0.074	-4615.76	0.014	0.064	-15759.87
$MPT(\Pi^{\mu,\sigma^2}, \mathcal{A}^c)$	2	0.101	0.309	-253.18	0.032	0.068	-626.07
	4	0.057	0.171	-448.85	0.020	0.038	-1876.51
	40	0.026	0.066	-4602.19	0.015	0.062	-15761.02

Table 2: SIMCE Dataset: Mean and standard deviation of the total number of correct answers (total score) by type of school in the SIMCE test.

Type of school	Number of examinees	Percentage	Mean	Standard Deviation
Public I (PUI)	16,200	14.7	20.48	7.51
Public II (PUII)	42,430	38.6	20.50	7.54
Private I (MI)	42,717	38.8	23.66	8.29
Private II (PR)	8,634	7.9	32.50	8.36
Full population	109,981	100.0	22.59	8.45

Table 3: SIMCE Dataset: Logarithm of the the Pseudo-Marginal Likelihood (LPML) and Pseudo-Bayes Factors (PBF) for the comparison of semiparametric versions of the model versus the parametric normal case.

Prior	LPML	$2\log$PBF
$N(\mu, \sigma^2)$	-26551.1	0.0
$MDP(M = 1, N(\mu, \sigma^2))$	-26534.6	33.1
$MDP(M, N(\mu, \sigma^2))$	-26526.1	49.9
$MPT(\Pi^{\mu,\sigma^2}, \mathcal{A}^1)$	-26521.6	59.1
$MPT(\Pi^{\mu,\sigma^2}, \mathcal{A}^c)$	-26521.8	58.7

Table 4: SIMCE Dataset: Bayesian posterior estimates for 5 randomly selected difficulty parameters under different nonparametric priors. The parametric normal case is also presented for comparison purposes.

Prior	Difficulty	Mean	Median	Std Dev	95% HPD Interval
$N(\mu, \sigma^2)$	β_2	-1.507	-1.502	0.302	(-2.106 ;-0.929)
	β_{10}	0.681	0.684	0.240	(0.207 ; 1.170)
	β_{12}	2.579	2.581	0.272	(2.032 ; 3.103)
	β_{29}	-0.203	-0.197	0.248	(-0.714 ; 0.256)
	β_{44}	2.612	2.615	0.263	(2.112 ; 3.146)
$MDP(M = 1, N(\mu, \sigma^2))$	β_2	-1.497	-1.493	0.298	(-2.082 ;-0.928)
	β_{10}	0.655	0.666	0.240	(0.167 ; 1.108)
	β_{12}	2.549	2.553	0.270	(2.017 ; 3.071)
	β_{29}	-0.219	-0.211	0.249	(-0.713 ; 0.260)
	β_{44}	2.594	2.598	0.267	(2.073 ; 3.107)
$MDP(M, N(\mu, \sigma^2))$	β_2	-1.512	-1.502	0.286	(-2.090 ;-0.977)
	β_{10}	0.645	0.643	0.215	(0.210 ; 1.055)
	β_{12}	2.538	2.533	0.245	(2.078 ; 3.027)
	β_{29}	-0.231	-0.229	0.225	(-0.676 ; 0.203)
	β_{44}	2.569	2.564	0.241	(2.105 ; 3.048)
$MPT(\Pi^{\mu,\sigma^2}, \mathcal{A}^1)$	β_2	-1.582	-1.580	0.284	(-2.154 ;-1.041)
	β_3	0.595	0.599	0.219	(0.159 ; 1.001)
	β_4	2.492	2.491	0.243	(2.026 ; 2.969)
	β_5	-0.289	-0.285	0.227	(-0.729 ; 0.156)
	β_6	2.520	2.516	0.246	(2.042 ; 3.001)
$MPT(\Pi^{\mu,\sigma^2}, \mathcal{A}^c)$	β_2	-1.551	-1.545	0.299	(-2.107 ;-0.944)
	β_3	0.618	0.616	0.227	(0.171 ; 1.048)
	β_4	2.519	2.517	0.255	(2.026 ; 3.010)
	β_5	-0.266	-0.270	0.238	(-0.721 ; 0.209)
	β_6	2.553	2.547	0.257	(2.059 ; 3.073)

Table 5: French Written Test Dataset: Logarithm of the the Pseudo-Marginal Likelihood (LPML) and Pseudo-Bayes Factors (PBF) for the comparison of semiparametric versions of the model versus the parametric normal case.

Prior	LPML	$2\log$PBF
$N(\mu, \sigma^2)$	-6610.300	0.0
$MDP(M = 1, N(\mu, \sigma^2))$	-6608.472	3.655
$MDP(M, N(\mu, \sigma^2))$	-6610.086	0.427
$MPT(\Pi^{\mu,\sigma^2}, \mathcal{A}^1)$	-6613.796	-4.989
$MPT(\Pi^{\mu,\sigma^2}, \mathcal{A}^c)$	-6599.572	21.455

Table 6: French Written Test Dataset: Bayesian posterior estimates for difficulty parameters under different nonparametric priors.

Prior	Difficulty	Mean	Median	Std Dev	95% HPD Interval
$N(\mu, \sigma^2)$	β_2	0.369	0.368	0.078	(0.216 ; 0.521)
	β_3	-0.574	-0.574	0.036	(-0.644 ;-0.502)
	β_4	0.325	0.325	0.042	(0.242 ; 0.407)
	β_5	0.677	0.677	0.034	(0.609 ; 0.742)
	β_6	0.065	0.066	0.035	(-0.005 ; 0.135)
$MDP(M = 1, N(\mu, \sigma^2))$	β_2	0.369	0.369	0.078	(0.223 ; 0.532)
	β_3	-0.574	-0.574	0.036	(-0.644 ;-0.502)
	β_4	0.325	0.325	0.043	(0.241 ; 0.409)
	β_5	0.677	0.677	0.034	(0.612 ; 0.744)
	β_6	0.065	0.064	0.036	(-0.005 ; 0.134)
$MDP(M, N(\mu, \sigma^2))$	β_2	0.370	0.369	0.077	(0.221 ; 0.522)
	β_3	-0.574	-0.573	0.036	(-0.646 ;-0.504)
	β_4	0.324	0.324	0.043	(0.240 ; 0.406)
	β_5	0.677	0.678	0.034	(0.609 ; 0.742)
	β_6	0.065	0.065	0.035	(-0.002 ; 0.133)
$MPT(\Pi^{\mu,\sigma^2}, \mathcal{A}^1)$	β_2	0.368	0.367	0.077	(0.218 ; 0.520)
	β_3	-0.574	-0.575	0.036	(-0.646 ;-0.503)
	β_4	0.325	0.324	0.042	(0.246 ; 0.410)
	β_5	0.677	0.677	0.034	(0.609 ; 0.742)
	β_6	0.065	0.065	0.035	(-0.004 ; 0.132)
$MPT(\Pi^{\mu,\sigma^2}, \mathcal{A}^c)$	β_2	0.372	0.371	0.078	(0.226 ; 0.531)
	β_3	-0.571	-0.571	0.037	(-0.643 ;-0.496)
	β_4	0.327	0.327	0.044	(0.243 ; 0.413)
	β_5	0.680	0.679	0.035	(0.610 ; 0.745)
	β_6	0.067	0.067	0.036	(-0.005 ; 0.136)