*The Annals of Applied Statistics*
2013, Vol. 7, No. 4, 2361–2383
DOI: 10.1214/13-AOAS675


# A SEMIPARAMETRIC APPROACH TO MIXED OUTCOME LATENT VARIABLE MODELS: ESTIMATING THE ASSOCIATION BETWEEN COGNITION AND REGIONAL BRAIN VOLUMES[1]

BY JONATHAN GRUHL*, ELENA A. EROSHEVA* AND PAUL K. CRANE†

*University of Washington** *and Harborview Medical Center*†

Multivariate data that combine binary, categorical, count and continuous outcomes are common in the social and health sciences. We propose a semiparametric Bayesian latent variable model for multivariate data of arbitrary type that does not require specification of conditional distributions. Drawing on the extended rank likelihood method by Hoff [*Ann. Appl. Stat.* **1** (2007) 265–283], we develop a semiparametric approach for latent variable modeling with mixed outcomes and propose associated Markov chain Monte Carlo estimation methods. Motivated by cognitive testing data, we focus on bifactor models, a special case of factor analysis. We employ our semiparametric Bayesian latent variable model to investigate the association between cognitive outcomes and MRI-measured regional brain volumes.

**1. Introduction.** Multivariate outcomes are common in medical and social studies. Latent variable models provide means for studying the interdependencies among multiple outcomes perceived as measures of a common concept or concepts. These outcomes in many cases may be of mixed types in the sense that some may be binary, others may be counts, and yet others may be continuous. Most common latent variable models, however, have been developed for outcomes of the same type. For example, standard factor analysis models [Bartholomew, Knott and Moustaki (2011)] assume normally distributed outcomes, item response theory models [van der Linden and Hambleton (1997)] are typically applied to binary responses, and graded response [Samejima (1969)] and generalized partial credit models [Muraki (1992)] have been developed specifically for ordered categorical data.

Existing research on latent variable models for mixed outcomes is largely focused on two parametric approaches. The first approach is to specify a different generalized linear model for each outcome that best suits its type (e.g., binary, count, ordered categorical) and to include shared latent variables as predictors that induce dependence among the outcomes. Sammel, Ryan and Legler (1997)

Received March 2012; revised July 2013.

[1]Supported in part by Grant R01 AG 029672 from the National Institute on Aging. Data collection was supported by Grant P01 AG12435.

*Key words and phrases.* Latent variable model, Bayesian hierarchical model, extended rank likelihood, cognitive outcomes.

and Moustaki and Knott (2000) developed this approach, referred to as generalized latent trait modeling, employing the EM algorithm for estimation. In a Bayesian framework, Dunson (2003) extended the generalized latent trait models to allow for repeated measurements, serial correlations in the latent variables and individual-specific response behavior. The second approach to analyzing mixed discrete and continuous outcomes with latent variables is the underlying latent response approach where observed mixed outcomes are assumed to have underlying latent responses that are continuous and normally distributed. Introduction of the continuous latent responses enables one to proceed with the analysis as one might for any multivariate normal data. To map the underlying latent responses to observed mixed outcomes, one must estimate threshold parameters. In this context, Shi and Lee (1998) employed Bayesian estimation for factor analysis with polytomous and continuous outcomes. However, as noted by Dunson (2003), the underlying latent response approach is limited in that some observed outcome types such as counts may not be easily linked to underlying continuous responses.

Generalized latent trait models can be extended to account for additional types of outcomes [Skrondal and Rabe-Hesketh (2004)], including censored and duration outcomes. However, accommodating many possible types of outcomes that one may encounter in practice may be time-consuming, susceptible to misspecification and of little interest in its own right.

Our motivating example is a data set from a large multicenter study called the Subcortical Ischemic Vascular Dementia (SIVD) Program Project Grant [Chui et al. (2006)]. The SIVD study collects serial imaging and neuropsychological data from a large group of study participants. One major study goal is to further elucidate relationships between brain structure (as measured by MRI imaging) and function (as measured by performance on neuropsychological tests). In particular, investigators were especially interested in cerebrovascular disease as manifested on MRI. Thus, we focus our analysis on one particular cognitive domain, namely, executive functioning, that is thought to be particularly susceptible to cerebrovascular disease [Hachinski et al. (2006)]. Executive functioning refers to higher order cognitive tasks ("executive" tasks) such as working memory, set shifting, inhibition and other frontal lobe-mediated functions. The SIVD study follows individuals longitudinally until death, collecting results from repeated neuropsychological testing and brain imaging. In this paper, we are concerned with relating individual levels of executive functioning at the initial SIVD study visit to the concurrent MRI-measured amount of white matter hyperintensities (WMH) located in the frontal lobe of the brain. Executive functioning capabilities may be particularly sensitive to white matter hyperintensities in this region [Kuczynski et al. (2010)].

The SIVD neuropsychological battery includes 21 distinct indicators that can be conceptualized as measuring some facet of executive functioning. We refer to the executive functioning-related outcomes as "indicators," as they include some elements that are scales by themselves and other elements that are subsets of scales. Observed responses to the SIVD neuropsychological tests items are diverse in their

types. In addition to binary and ordered categorical outcomes, the SIVD neuropsychological indicators include count as well as censored count data.

In this paper, we develop a semiparametric approach to mixed outcome latent variable models that avoids specification of outcome conditional distributions given the latent variables. Following the extended rank likelihood approach of Hoff (2007), we start by assuming the existence of continuous latent responses underlying each observed outcome. We then make use of the fact that the ordering of the underlying latent responses is assumed to be consistent with the ordering of the observed outcomes. This approach is similar to that of Shi and Lee (1998) but does not require estimating unknown thresholds. When the data are continuous, our approach is analogous to the use of a rank likelihood [Pettitt (1982)]. When the data are discrete, our approach relies on the assumption that the ordering of the latent responses is consistent with the partial ordering of the observed outcomes. Hoff (2007) introduced this general approach for estimating parameters of a semiparametric Gaussian copula model with arbitrary marginal distributions and designated the resulting likelihood as the *extended rank likelihood*. Dobra and Lenkoski (2011) applied the extended rank likelihood methods to the estimation of graphical models for multivariate mixed outcomes.

Motivated by SIVD cognitive testing data, we specify a bifactor latent structure for the semiparametric latent variable model. The bifactor structure assumes existence of a general factor and some secondary factors that account for residual dependency among groups of items [Holzinger and Swineford (1937), Reise, Morizot and Hays (2007)]. The bifactor model is a useful tool for modeling the neuropsychological battery used in the SIVD study, as it retains a single underlying executive functioning factor while accounting for local dependencies among groups of related items. The original idea for this work was presented earlier by [Gruhl, Erosheva and Crane (2010, 2011)]. Murray et al. (2013) recently proposed a closely related factor analytic model for mixed data.

The remainder of this paper is organized as follows. We review the semiparametric Gaussian copula model and introduce the new semiparametric latent variable model in Section 2. We develop Bayesian estimation approaches for the semiparametric latent variable model in Section 3. We extend the model hierarchically to include covariates in Section 4. In Section 5 we briefly demonstrate the performance of the model using simulated data before focusing on the analysis of the SIVD data.

## 2. Semiparametric latent variable model for mixed outcomes.

2.1. *Model formulation.* Let $i = 1, \ldots, I$ denote the $i$th participant, and let $j = 1, \ldots, J$ denote the $j$th outcome. Let $y_{ij}$ denote the observed response of participant $i$ on outcome $j$ with marginal distribution $F_j$, then $y_{ij}$ can be represented as $y_{ij} = F_j^{-1}(u_{ij})$, where $u_{ij}$ is a uniform $(0, 1)$ random variable. An equivalent

representation is $y_{ij} = F_j^{-1}[\Phi(z_{ij})]$, where $\Phi(\cdot)$ denotes the normal CDF and $z_{ij}$ is distributed standard normal. The unobserved variables $z_{ij}$ are latent responses underlying each observed response $y_{ij}$. Assuming that the correlation of $z_{ij}$ with $z_{ij'}$, $1 \le j < j' \le J$, is specified by the $J \times J$ correlation matrix $\mathbf{C}$, the Gaussian copula model is

$$\mathbf{z}_1, \ldots, \mathbf{z}_n | \mathbf{C} \sim \text{i.i.d. } \mathrm{N}(\mathbf{0}, \mathbf{C}), \tag{1}$$

$$y_{ij} = F_j^{-1}\big[\Phi(z_{ij})\big]. \tag{2}$$

Here, $\mathbf{z}_i$ is the $J$-length vector of latent responses $z_{ij}$ for participant $i$.

In some analyses, the primary focus is on the estimation of the correlation matrix $\mathbf{C}$ and not the estimation of the marginal distributions $F_1, \ldots, F_J$. If the latent responses $z_{ij}$ were known, estimation of $\mathbf{C}$ could proceed using standard methods. Although the latent responses are unknown, Hoff (2007) noted that we do have rank information about the latent responses through the observed responses because $y_{ij} < y_{ij'}$ implies $z_{ij} < z_{ij'}$. If we denote the full set of latent responses by $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_I)^T$ and the full set of observed responses by $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_I)^T$, then $\mathbf{Z} \in D(\mathbf{Y})$, where

$$D(\mathbf{Y}) = \Big\{ \mathbf{Z} \in \mathbb{R}^{I \times J} : \max_k \{z_{kj} : y_{kj} < y_{ij}\} < z_{ij} < \min_k \{z_{kj} : y_{ij} < y_{kj}\} \ \forall i, j \Big\}. \tag{3}$$

One can then construct a likelihood for $\mathbf{C}$ that does not depend on the specification of the marginal distributions $F_1, \ldots, F_J$ by focusing on the probability of the event, $\mathbf{Z} \in D(\mathbf{Y})$:

$$\begin{aligned} \Pr\big(\mathbf{Z} \in D(\mathbf{Y}) | \mathbf{C}, F_1, \ldots, F_J\big) &= \int_{D(\mathbf{Y})} p(\mathbf{Z}|\mathbf{C})\, d\mathbf{Z} \\ &= \Pr\big(\mathbf{Z} \in D(\mathbf{Y}) | \mathbf{C}\big). \end{aligned} \tag{4}$$

Equation (3) enables the following decomposition of the density of $\mathbf{Y}$:

$$\begin{aligned} &p(\mathbf{Y}|\mathbf{C}, F_1, \ldots, F_J) \\ &\quad = p\big(\mathbf{Y}, \mathbf{Z} \in D(\mathbf{Y}) | \mathbf{C}, F_1, \ldots, F_J\big) \\ &\quad = \Pr\big(\mathbf{Z} \in D(\mathbf{Y}) | \mathbf{C}, F_1, \ldots, F_J\big) \times p\big(\mathbf{Y} | \mathbf{Z} \in D(\mathbf{Y}), \mathbf{C}, F_1, \ldots, F_J\big) \\ &\quad = \Pr\big(\mathbf{Z} \in D(\mathbf{Y}) | \mathbf{C}\big) \times p\big(\mathbf{Y} | \mathbf{Z} \in D(\mathbf{Y}), \mathbf{C}, F_1, \ldots, F_J\big). \end{aligned}$$

This decomposition uses the fact that the probability of the event $\mathbf{Z} \in D(\mathbf{Y})$ does not depend on the marginal distributions $F_1, \ldots, F_J$ and that the event $\mathbf{Z} \in D(\mathbf{Y})$ occurs whenever $\mathbf{Y}$ is observed. By using $\Pr(\mathbf{Z} \in D(\mathbf{Y})|\mathbf{C})$ as the likelihood function, the dependence structure of $\mathbf{Y}$ can be estimated through $\mathbf{C}$ without any knowledge or assumptions about the marginal distributions. More details on the semiparametric Gaussian copula model can be found in Hoff (2007, 2009).

In the context of latent variable modeling, the main interest is not just in estimating the correlations among observed variables $\mathbf{C}$ but in characterizing the interdependencies in multivariate observed responses through a latent variable model. Latent variable models, in turn, place constraints on the matrix of correlations among the observed responses and seek a more parsimonious description of the dependence structure. Factor analysis is the most common type of latent variable model with continuous latent variables and continuous outcomes.

To develop a semiparametric approach for factor analysis with mixed outcomes, assume $Q$ factors, let $\boldsymbol{\eta}_i$ be a vector of factor scores for individual $i$ and $\mathbf{H} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_I)^T$ be the $I \times Q$ factor matrix. Let $\Lambda$ denote the $J \times Q$ matrix of factor loadings. We define our semiparametric latent variable model as

$$\boldsymbol{\eta}_i \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_Q), \tag{5}$$

$$\mathbf{z}_i | \Lambda, \boldsymbol{\eta}_i \sim \mathrm{N}(\Lambda \boldsymbol{\eta}_i, \mathbf{I}_J), \tag{6}$$

$$y_{ij} = g_j(z_{ij}). \tag{7}$$

Here, we define $g_j(z_{ij}) = F_j^{-1}(\Phi[z_{ij}/\sqrt{1 + \boldsymbol{\lambda}_j^T \boldsymbol{\lambda}_j}])$, where $\boldsymbol{\lambda}_j$ denotes the $j$th row of $\boldsymbol{\Lambda}$ and the marginal distribution $F_j$ remains unspecified. Note that the functions $g_j(\cdot)$ are nondecreasing and preserve the orderings. The model given by equations (5)–(7) does not rely on the unrestricted correlation matrix $\mathbf{C}$ as does the Gaussian copula model. Assuming that a factor analytic model is appropriate for the data, it constrains the dependencies among the elements of $\mathbf{z}_i$ to be consistent with the functional form of $\mathbf{I}_J + \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T$. As a result, the proposed semiparametric latent variable model is a structured case of the semiparametric Gaussian copula model and can be viewed as a semiparametric form of copula structure analysis [Klüppelberg and Kuhn (2009)].

The general framework of the semiparametric latent variable model given by equations (5)–(7) can be used for any special cases of factor analysis. In this paper, motivated by the substantive background information on the SIVD cognitive testing data, we focus on bifactor models. We define bifactor models as having a specific structure on the loading matrix, $\boldsymbol{\Lambda}$, where each outcome loads on the primary factor and may load on one or more of the secondary factors [Reise, Morizot and Hays (2007)]. Most commonly, bifactor models are applied such that an outcome loads on at most one secondary factor.

2.2. *Model identification*. The lack of identifiability of factor analysis models that is due to rotational invariance is well known [Anderson (2003), Dunn (1973), Jennrich (1978), Jöreskog (1969)]. If we define new factor loadings and new factor scores by $\tilde{\boldsymbol{\Lambda}} = \boldsymbol{\Lambda}\mathbf{T}$ and $\tilde{\boldsymbol{\eta}}_i = \mathbf{T}^{-1}\boldsymbol{\eta}_i$, where $\mathbf{T}$ is an orthonormal $Q \times Q$ matrix, then the model

$$\mathbf{z}_i | \boldsymbol{\Lambda}, \boldsymbol{\eta}_i \sim \mathrm{N}(\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\eta}}_i, \mathbf{I}_J) \tag{8}$$

is indistinguishable from the model in equation (6). In the case where the covariance of $\boldsymbol{\eta}_i$ is not restricted to the identity matrix, any nonsingular $Q \times Q$ matrix $\mathbf{T}$ results in the same indeterminacy. In this more general case, we must place $Q^2$ constraints to prevent this rotational invariance. When we restrict the covariance of $\boldsymbol{\eta}_i$ to the identity matrix, this restriction places $\frac{1}{2}Q(Q+1)$ constraints on the model. We are then left with $\frac{1}{2}Q(Q-1)$ additional constraints to place on the model. We may satisfy this requirement by assuming a bifactor structure with $\frac{1}{2}Q(Q-1)$ zeros in the matrix of loadings $\boldsymbol{\Lambda}$ [Anderson (2003)]. While these restrictions may resolve rotational invariance, the issue of reflection invariance typically remains [Dunn (1973), Jennrich (1978)]. Reflection invariance results from the the fact that the signs of the loadings in any column in $\boldsymbol{\Lambda}$ may be switched. Thus, if $\mathbf{D}$ is a diagonal matrix of 1's and $-1$'s precipitating the sign changes, $\mathbf{H}\boldsymbol{\Lambda}^T = \mathbf{HDD}\boldsymbol{\Lambda}^T = \tilde{\mathbf{H}}\tilde{\boldsymbol{\Lambda}}^T$.

Geweke and Zhou (1996) proposed an approach that addresses identifiability of factor models by constraining all upper diagonal elements in the matrix of factor loadings to zero and requiring all diagonal elements to be positive. This approach has been used successfully in Bayesian exploratory factor analysis [Ghosh and Dunson (2008, 2009), Lopes and West (2004)], but cannot be used with bifactor models because the placement of structural zeros in most cases will be incompatible with fixing all upper diagonal elements of the matrix of loadings to zero. Congdon (2003) and Congdon (2006) suggested the use of a prior that would place additional constraints on the signs of some of the factor loadings to resolve the issue of reflection invariance. However, it has been shown that different choices of parameters for constraint placement could have a serious impact on model fit in complex factor models [Millsap (2001)]. Thus, in our work, we rely on the relabeling algorithm proposed by Erosheva and Curtis (2011) to resolve reflection invariance. This algorithm relies on a decision-theoretic approach and resolves the sign-switching behavior in Bayesian factor analysis in a similar fashion to the relabeling algorithm introduced to address the label-switching problem in mixture models [Stephens (2000)]. It does not require making preferential choices among variables for constraint placement.

In the semiparametric latent variable model, unlike the standard factor analysis model, specific means and variances are not identifiable. Let

$$\tilde{z}_{ij} = \mu_j + \sigma_j z_{ij}, \tag{9}$$

where $\mu_j$ and $\sigma_j$ are the specific mean and variance for item $j$. Moreover, if $\tilde{\mathbf{Z}} \in \mathbb{R}^{I \times J}$ denotes the matrix of elements $\tilde{z}_{ij}$ and

$$\tilde{D}(\mathbf{Y}) = \{\tilde{\mathbf{Z}} : \max\{\tilde{z}_{kj} : y_{kj} < y_{ij}\} < \tilde{z}_{ij} < \min\{\tilde{z}_{kj} : y_{ij} < y_{kj}\}\}, \tag{10}$$

then

$$\Pr(\tilde{\mathbf{Z}} \in \tilde{D}(\mathbf{Y}) | \boldsymbol{\Lambda}, \mathbf{H}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \Pr(\mathbf{Z} \in D(\mathbf{Y}) | \boldsymbol{\Lambda}, \mathbf{H}). \tag{11}$$

Thus, shifts in location and scale of the latent responses will not alter the probability of belonging to the set of feasible latent response values implied by orderings of the observed responses. As such, we set the specific means at $\boldsymbol{\mu} = 0$ and the specific variances at $\boldsymbol{\Sigma} = \mathbf{I}_J$.

**3. Estimation.** We employ a parameter expansion approach [Liu, Rubin and Wu (1998), Liu and Wu (1999)] for Markov chain Monte Carlo (MCMC) sampling, following the work of Ghosh and Dunson (2009) on efficient computation for Bayesian factor analysis. We found that this method outperforms a Gibbs sampling algorithm with standard semi-conjugate priors for factor analysis [Ghosh and Dunson (2009), Shi and Lee (1998)] in that it reduces autocorrelation among the MCMC draws and results in greater effective sample sizes.

3.1. *Parameter expansion approach.* The central idea behind the parameter expansion approach, using the terminology of Ghosh and Dunson (2009), is to start with a working model that is an overparameterized version of the initial inferential model. After proceeding through MCMC sampling, a transformation is used to relate the draws from the working model to draws from the inferential model. For our application, the overparameterized model is

$$\mathbf{z}_i^* \sim \mathrm{N}\big(\boldsymbol{\Lambda}^* \boldsymbol{\eta}_i^*, \boldsymbol{\Sigma}\big), \tag{12}$$

$$\boldsymbol{\eta}_i^* \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Psi}), \tag{13}$$

where $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$ are diagonal matrices that are no longer restricted to identity matrices. The latent responses $\mathbf{z}_i^*$, the latent variables $\boldsymbol{\eta}_i^*$ and the loadings $\boldsymbol{\Lambda}^*$ are unidentified in this working model. The transformations from the working model to the inferential model are then specified as

$$\begin{aligned} \boldsymbol{\eta}_i &= \boldsymbol{\Psi}^{-1/2} \boldsymbol{\eta}_i^*, \\ \mathbf{z}_i &= \boldsymbol{\Sigma}^{-1/2} \mathbf{z}_i^*, \\ \boldsymbol{\Lambda} &= \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Lambda}^* \boldsymbol{\Psi}^{1/2}. \end{aligned} \tag{14}$$

To sample from the working model, we must specify priors for the diagonal elements of $\boldsymbol{\Psi}$ and $\boldsymbol{\Sigma}$ as well as for $\boldsymbol{\Lambda}^*$. We specify these priors in terms of the precisions $\psi_q^{-2}$ and $\sigma_j^{-2}$. In addition, we denote by $\boldsymbol{\lambda}_j^{*\prime}$ the nonzero elements of the $j$th row of $\boldsymbol{\Lambda}^*$. The prior on $\boldsymbol{\lambda}_j'$ is then induced through the priors on $\psi_q^{-2}$, $\sigma_j^{-2}$ and $\boldsymbol{\lambda}_j^{*\prime}$, rather than being specified directly. Our priors are

$$\begin{aligned} \psi_q^{-2} &\sim \mathrm{Gamma}(\phi_\psi, \nu_\psi), \\ \sigma_j^{-2} &\sim \mathrm{Gamma}(\phi_\sigma, \nu_\sigma), \\ \boldsymbol{\lambda}_j^{*\prime} &\sim \mathrm{N}(\mathbf{m}_{\lambda_j^{*\prime}}, \mathbf{S}_{\lambda_j^{*\prime}}). \end{aligned} \tag{15}$$

The structural zeros in the matrix of loadings $\mathbf{\Lambda}$ are specified in accordance with our substantive understanding of the research problem at hand. However, we must have enough zeros so that the model can be identified since we rely on the placement of these structural zeros to resolve rotational invariance [Dunn (1973), Jennrich (1978), Jöreskog (1969), Loken (2005)]. Formally, we specify the prior for these structural zero elements as

$$\lambda_{jq}^{*} \sim \delta_0, \tag{16}$$

where $\delta_0$ is a distribution with its point mass concentrated at 0. We estimate loadings with no additional constraints on their signs. As discussed in Section 2, we then deal with potential multiple modes of the posterior that are due to reflection invariance by applying the relabeling algorithm proposed by Erosheva and Curtis (2011).

We now develop the parameter-expanded Gibbs algorithm for sampling factors $\mathbf{H}$ and loadings $\mathbf{\Lambda}$. Because the extended rank likelihood $\Pr(\mathbf{Z}^* \in D(\mathbf{Y})|\mathbf{\Lambda}^*, \mathbf{H}^*, \mathbf{\Sigma})$ involves a complicated integral, any expressions involving it will be difficult to compute. We avoid having to compute this integral by drawing from the joint posterior of $(\mathbf{Z}^*, \mathbf{H}^*, \mathbf{\Lambda}^*, \mathbf{\Sigma}, \mathbf{\Psi})$ via Gibbs sampling. Given $\mathbf{Z}^* = \mathbf{z}^*$ and $\mathbf{Z}^* \in D(\mathbf{Y})$, the full conditional density of $\mathbf{\Lambda}^*$ can be written as

$$p\big(\mathbf{\Lambda}^*|\mathbf{H}^*, \mathbf{Z}^* = \mathbf{z}^*, \mathbf{Z}^* \in D(\mathbf{Y}), \mathbf{\Sigma}\big) = p\big(\mathbf{\Lambda}^*|\mathbf{H}^*, \mathbf{Z}^* = \mathbf{z}^*, \mathbf{\Sigma}\big)$$

because the current draw values $\mathbf{Z}^* = \mathbf{z}^*$ imply $\mathbf{Z}^* \in D(\mathbf{Y})$. A similar simplification may be made with the full conditional density of $\mathbf{H}^*$. Given $\mathbf{\Lambda}^*, \mathbf{H}^*, \mathbf{Z}^* \in D(\mathbf{Y})$, $\mathbf{\Sigma}$ and $\mathbf{Z}^*_{(-i)(-j)}$, the full conditional density of $z_{ij}$ is proportional to a normal density with mean $(\boldsymbol{\lambda}_j^*)^T \boldsymbol{\eta}_i^*$ and variance $\sigma_j^2$ that is restricted to the region specified by $D(\mathbf{Y})$. Our Gibbs sampling procedure for the working model proceeds according to the following steps:

1. *Draw latent responses* $\mathbf{Z}^*$. For each $i$ and $j$, sample $z_{ij}^*$ from a truncated normal distribution according to

$$z_{ij}^* \sim \mathrm{TN}_{(z_l^*, z_u^*)}\big((\boldsymbol{\lambda}_j^*)^T \boldsymbol{\eta}_i^*, \sigma_j^2\big), \tag{17}$$

where TN denotes truncated normal and $z_l^*, z_u^*$ define the lower and upper truncation points:

$$z_l^* = \max_k \big\{z_{kj}^* : y_{kj} < y_{ij}\big\}, \tag{18}$$

$$z_u^* = \min_k \big\{z_{kj}^* : y_{kj} > y_{ij}\big\}. \tag{19}$$

2. *Draw latent variables* $\mathbf{H}^*$. For each $i$, draw directly from the full conditional distribution for $\boldsymbol{\eta}_i^*$ as follows:

$$\begin{aligned} \eta_i^* \sim \mathrm{N}\big(&\big(\mathbf{\Psi}^{-1} + (\mathbf{\Lambda}^*)^T \mathbf{\Sigma}^{-1} \mathbf{\Lambda}^*\big)^{-1} (\mathbf{\Lambda}^*)^T \Sigma^{-1} \mathbf{z}_i^*, \\ &\big(\mathbf{\Psi}^{-1} + (\mathbf{\Lambda}^*)^T \mathbf{\Sigma}^{-1} \mathbf{\Lambda}^*\big)^{-1}\big). \end{aligned} \tag{20}$$

3. *Draw loadings* $\boldsymbol{\Lambda}^*$. For each $j$, draw from the full conditional distribution for the nonzero loadings $\boldsymbol{\lambda}_j^{*\prime}$:

$$\begin{aligned}\boldsymbol{\lambda}_j^{*\prime} \sim \mathrm{N}\big((\mathbf{S}_{\lambda_j'}^{-1} + \sigma_j^{-2}(\mathbf{H}_j^{*\prime})^T\mathbf{H}_j^{*\prime})^{-1}(\mathbf{S}_{\lambda_j'}^{-1}m_{\lambda_j'} + \sigma_j^{-2}(\mathbf{H}_j^{*\prime})^T\mathbf{z}_j^*),\\ (\mathbf{S}_{\lambda_j'}^{-1} + \sigma_j^{-2}(\mathbf{H}_j^{*\prime})^T\mathbf{H}_j^{*\prime})^{-1}\big),\end{aligned} \tag{21}$$

where $\mathbf{H}_j^{*\prime}$ is a matrix comprised of the columns of $\mathbf{H}^*$ for which there are nonzero loadings in $\boldsymbol{\lambda}_j$.

4. *Draw inverse covariance matrix* $\boldsymbol{\Psi}^{-1}$. For each $q$, draw the diagonal element $\psi_q^{-2}$ of $\boldsymbol{\Psi}^{-1}$ from the full conditional distribution:

$$\psi_q^{-2} \sim \mathrm{Gamma}\Big(\phi_\psi + I/2, \nu_\psi + \frac{1}{2}\sum_i \eta_{iq}^2\Big), \tag{22}$$

where $I$ is the number of participants.

5. *Draw inverse covariance matrix* $\boldsymbol{\Sigma}^{-1}$. For each $j$, draw the diagonal element $\sigma_j^{-2}$ of $\boldsymbol{\Sigma}^{-1}$ from the full conditional distribution:

$$\sigma_j^{-2} \sim \mathrm{Gamma}\big(\phi_\sigma + I/2, \nu_\sigma + \tfrac{1}{2}(\mathbf{z}_j - \mathbf{H}\boldsymbol{\lambda}_j)^T(\mathbf{z}_j - \mathbf{H}\boldsymbol{\lambda}_j)\big). \tag{23}$$

After discarding some number of initial draws as burn-in, we transform the remaining draws using equations (14) as part of a postprocessing step to obtain posterior draws from our inferential model. The only remaining steps are to apply the relabeling algorithm of Erosheva and Curtis (2011), assess convergence and calculate posterior summaries for the parameters in the inferential model.

Our application of parameter expansion to factor analysis models induces prior distributions that are different from standard semi-conjugate priors in factor analysis. If the prior covariance matrix on $\boldsymbol{\lambda}_j'$ is diagonal, the prior induced on $\lambda_{jq}'$ by the parameter expansion is the product of the normal distribution prior on $\lambda_{jq}^{*\prime}$ and the square root of a ratio of gamma distribution priors on $\sigma_j^{-2}$ and $\psi_q^{-2}$. The ratio of gamma distributed random variables has a compound gamma distribution which is a form of the generalized beta prime distribution with the shape parameter fixed to 1. If we integrate out this ratio, the prior for $\lambda_{jq}$ is a scale mixture of normals [West (1987)] with a compound gamma mixing density.

The induced prior on the matrix $\boldsymbol{\Lambda}$ results in correlations among elements of the same column and elements of the same row. As discussed in Ghosh and Dunson (2009), prior dependence in the factor loadings for the $q$th factor will result from the shared parameter $\psi_q^2$ in their respective prior distributions in the parameter expanded formulation. Similarly, the shared parameter $\sigma_j^2$ in the prior distributions for the factor loadings related to the $j$th outcome in the parameter expanded approach will induce prior dependence across rows of the factor loadings matrix.

In both the simulation and applied settings considered below, sensitivity analyses demonstrated that posterior estimates did not change meaningfully for various hyperparmater values for $\sigma_j^{-2}$ and $\psi_q^{-2}$. For values of $\nu_\sigma = 1, \phi_\sigma = 2, \nu_\psi = 1/2, \phi_\psi = 2$, the induced prior on the nonzero elements of $\mathbf{\Lambda}$ will have mean, variance and 2.5% and 97.5% quantiles close to that of a standard normal distribution. In their comparable model, Murray et al. (2013) use a shrinkage prior, explore its properties and develop a parameter-expanded approach with optimality properties.

3.2. *Generating replicated data for posterior predictive model checks*. Following Hoff (2007), we obtain posterior predictive distributions that incorporate uncertainty in estimation of the univariate marginal distributions. Let the superscript $(m)$ denote the $m$th replicate from the $m$th posterior draw of the parameter. We generate a new vector of latent responses, $z_{I+1}^{(m)}$, in addition to $I$ sets drawn as part of the Gibbs sampling algorithm, according to

$$\mathbf{z}_{I+1}^{(m)} \sim \mathrm{N}\big(\mathbf{0}, \mathbf{I}_J + \mathbf{\Lambda}^{(m)}\big(\mathbf{\Lambda}^{(m)}\big)^T\big). \tag{24}$$

If $z_{(I+1)j}^{(m)}$ falls between two latent responses, $z_{ij}^{(m)}$ and $z_{i'j}^{(m)}$, that share the same value on the original data scale (i.e., $y_{ij} = y_{i'j}$), then $y_{(I+1)j}^{(m)}$ must also take this value as $g_j(\cdot)$ is monotonic. If $z_{(I+1)j}^{(m)}$ falls between two latent responses, $z_{ij}^{(m)}$ and $z_{i'j}^{(m)}$, that do not share the same value on the original data scale, then we select the value, $y_{ij}$ or $y_{i'j}$, corresponding to the latent response to which $z_{(I+1)j}^{(m)}$ is closest. In the case of continuous observed responses, we use linear interpolation to obtain a value for $y_{(I+1)j}^{(m)}$.

**4. Hierarchical semiparametric latent variable model.** To relate covariates of interest to the primary factor, we extend the proposed model hierarchically. Previously, we assumed that

$$\boldsymbol{\eta}_i \sim \mathrm{N}(\mathbf{m}_{\eta_i}, \mathbf{\Psi}), \tag{25}$$

where $\mathbf{m}_{\eta_i} = \mathbf{0}$, $\mathbf{\Psi} = \mathbf{I}_Q$. We now replace the first element of $\mathbf{m}_{\eta_i}$ with a function of the covariates of interest denoted by the $P$-length vector $\mathbf{x}_i$:

$$\mathbf{m}_{\eta_i} = \big(x_i^T \boldsymbol{\beta}, 0, \ldots, 0\big)^T. \tag{26}$$

When we employ a parameter expansion approach for estimation,

$$\boldsymbol{\eta}_i^* \sim \mathrm{N}(\mathbf{m}_{\eta_i}, \mathbf{\Psi}), \tag{27}$$

$$\mathbf{m}_{\eta_i} = \big(x_i^T \boldsymbol{\beta}^*, 0, \ldots, 0\big)^T, \tag{28}$$

where the diagonal elements of $\mathbf{\Psi}$ are no longer restricted during MCMC. For $\boldsymbol{\beta}^*$, we specify the semi-conjugate prior:

$$\boldsymbol{\beta}^* \sim \mathrm{N}(\mathbf{m}_\beta, \mathbf{S}_\beta). \tag{29}$$

Moreover, to further facilitate efficient computation, we add an additional working parameter, $\alpha$, as suggested by Ghosh and Dunson (2009), so that

$$\mathbf{m}_{\eta_i} = (\alpha + x_i^T \boldsymbol{\beta}^*, 0, \ldots, 0)^T. \tag{30}$$

Relaxing the restriction on the mean of the latent variable promotes better mixing of the regression coefficients. For $\alpha$, we use the semi-conjugate prior:

$$\alpha \sim \mathrm{N}(m_\alpha, s_\alpha^2). \tag{31}$$

To estimate the hierarchical model [equations (25), (26)], we modify the steps for drawing $\boldsymbol{\eta}_i^*$ and $\boldsymbol{\Psi}$ in the sampling algorithm from Section 3 to account for the inclusion of covariates and the additional working parameter, $\alpha$, in $\mathbf{m}_{\eta_i}$. We sample $\boldsymbol{\beta}^*$ and $\alpha$ according to their full conditionals:

$$\begin{aligned}\boldsymbol{\beta}^* \sim \mathrm{N}\big(&(\psi_1^{-2}\mathbf{X}^T\mathbf{X} + \mathbf{S}_\beta^{-1})^{-1}(\psi_1^{-2}\mathbf{X}^T(\boldsymbol{\eta}_{q=1}^* - \mathbf{1}_I\alpha) + \mathbf{S}_\beta^{-1}\mathbf{m}_\beta),\\ &(\psi_1^{-2}\mathbf{X}^T\mathbf{X} + \mathbf{S}_\beta^{-1})^{-1}\big),\end{aligned} \tag{32}$$

$$\begin{aligned}\alpha \sim \mathrm{N}\big(&(\psi_1^{-2}I + s_\alpha^{-2})^{-1}(\psi_1^{-2}\mathbf{1}_I^T(\boldsymbol{\eta}_{q=1}^* - \mathbf{X}\boldsymbol{\beta}^*) + s_\alpha^{-2}m_\alpha),\\ &(\psi_1^{-2}I + s_\alpha^{-2})^{-1}\big),\end{aligned} \tag{33}$$

where $\mathbf{X}$ is an $I \times P$ matrix of covariates and $\boldsymbol{\eta}_{q=1}^*$ is a vector of the primary factor scores, the first column of $\mathbf{H}^*$. In the postprocessing stage for the parameter expansion approach, we make the transformations:

$$\boldsymbol{\eta}_i = \boldsymbol{\Psi}^{-1/2}(\boldsymbol{\eta}_i^* - \boldsymbol{\alpha}), \tag{34}$$

$$\boldsymbol{\beta} = \boldsymbol{\beta}^*\psi_1^{-1}, \tag{35}$$

where $\boldsymbol{\alpha} = (\alpha, 0, \ldots, 0)^T$.

## 5. Simulation data example and application to SIVD data.

5.1. *Simulated data example.* To test the semiparametric latent variable model, we examined the model's ability to recover data generating parameters using simulated data for $I = 500$ individuals on $J = 15$ outcomes. We applied three estimation approaches: the parameter expansion approach from Section 3, a variation of the parameter expansion approach where only the diagonal elements of $\boldsymbol{\Psi}$ are unrestricted during estimation, and a more standard Gibbs sampling approach [Ghosh and Dunson (2009), Shi and Lee (1998)]. The data generating process assumed a bifactor form with two secondary factors in addition to the primary factor. All outcomes loaded on the general factor; outcomes 1, 13, 14 and 15 loaded on one secondary factor; and outcomes 3, 4, 6 and 8 loaded on the other secondary factor. For each individual, we simulated $\boldsymbol{\eta}_i \sim \mathrm{N}(\mathbf{0}, \mathbf{I}_Q)$. We subsequently generated a matrix of latent responses, $\mathbf{Z}$, with mean $\mathbf{H}^T\boldsymbol{\Lambda}$. Finally, we randomly drew

cutoffs for each outcome in order to produce discretized "observed" responses from the continuous latent responses so that the number of unique values for each outcome ranged from 2 (outcome 1) to 30 (outcome 9).

To fit the model to the simulated data, we employed each estimation approach to generate 50,000 MCMC draws, the first 10,000 of which we discarded as burn-in. In addition, we thinned the posterior samples, keeping only every 10th draw. All estimation approaches did a good job of recovering the data generating values, but the parameter expansion estimation approach described in Section 3 displayed better mixing, less autocorrelation and larger effective sample sizes, sometimes by a factor of ten or more compared to a standard Gibbs sampling approach. We note that because the target distributions are not the same for the three approaches, measures such as effective sample size are not directly comparable. However, as in Ghosh and Dunson (2009), we use these measures as indicators of the quality of mixing using the different approaches. Finally, as discussed in Section 3, different choices for hyperparameter values did not result in meaningful differences in the posterior estimates.

5.2. *Application to SIVD data*. Participants were recruited to fill six broad groups in the Subcortical Ischemic Vascular Dementia (SIVD) study [Chui (2007)], comprised by three levels of cognitive functioning and two levels (absence vs. presence) of subcortical lacunes. Lacunes are small areas of dead brain tissue caused by blocked or restricted blood supply. The three levels of cognitive functioning groups were normal, mildly impaired and demented as determined by the Clinical Dementia Rating total score, a numerical rating that is based on medical history and clinical examination as well as other forms of assessment [Morris (1993, 1997)]. Among the data collected by SIVD are neuropsychological test results and standardized magnetic resonance imaging (MRI) scans of the participants' brains [Mungas et al. (2005)]. A computerized segmentation algorithm classified pixels from the MRI scans into different components, including white matter hyperintensities [Cardenas et al. (2001)].

We are interested in relating the individuals' level of executive functioning to the white matter hyperintensity volume located in the frontal lobe at individuals' first visit. White matter hyperintensities are areas of increased signal intensity that are commonly associated with older age. Among the outcomes available at first visit, we identified 21 indicators of executive functioning in the SIVD neuropsychological battery. These items included Digit Span, Visual Span, Verbal Fluency, Stroop Test and Mattis Dementia Rating Scale (MDRS) test items. We excluded MDRS outcomes M and N, as everyone except one participant received full credit on these outcomes. As a result, we used 19 of the 21 executive functioning outcomes in our analysis.

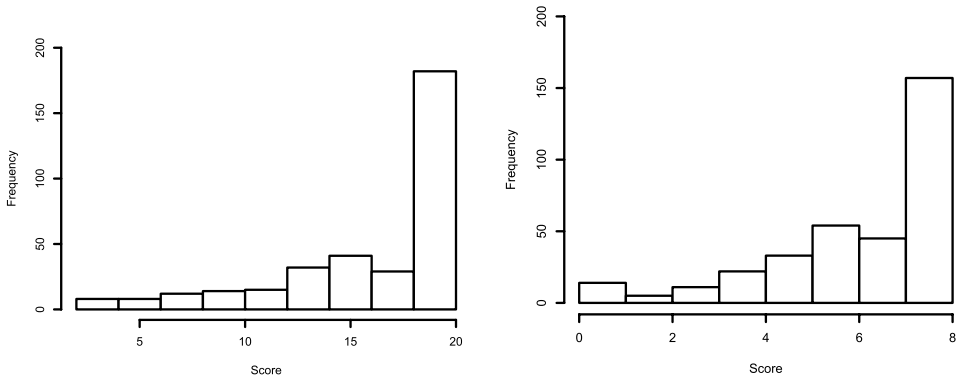
Table 1 displays basic information for the 19 outcomes as well as some summary statistics observed in the data for $I = 341$ participants. For this analysis, we considered only participants with a complete set of responses to the 19 executive

TABLE 1
*Summary statistics for $I = 341$ responses to 19 SIVD executive functioning outcomes as well as outcome type assignment. "RC Count" denotes a right-censored count outcome*

| | **Range** | **Mean** | **Median** | **Outcome type** |
|---|---|---|---|---|
| Digit Span Forward | 3–12 | 7.69 | 8 | Count |
| Digit Span Backward | 1–12 | 5.97 | 6 | Count |
| Visual Span Forward | 0–13 | 7.15 | 7 | Count |
| Visual Span Backward | 0–12 | 6.18 | 6 | Count |
| Verbal Fluency Letter F | 1–26 | 11.8 | 12 | Count |
| Verbal Fluency Letter A | 0–40 | 10.2 | 10 | Count |
| Verbal Fluency Letter S | 0–50 | 12.4 | 12 | Count |
| MDRS E | 2–20 | 16.64 | 19 | RC Count |
| MDRS G | 0–1 | 0.96 | 1 | Binary |
| MDRS H | 0–1 | 0.98 | 1 | Binary |
| MDRS I | 0–1 | 0.95 | 1 | Binary |
| MDRS J | 0–1 | 0.97 | 1 | Binary |
| MDRS K | 0–1 | 0.98 | 1 | Binary |
| MDRS L | 0–1 | 0.79 | 1 | Binary |
| MDRS O | 0–1 | 0.94 | 1 | Binary |
| MDRS V | 9–16 | 14.9 | 16 | RC Count |
| MDRS W | 0–8 | 6.44 | 7 | Ordered Cat. |
| MDRS X | 0–3 | 2.66 | 3 | Ordered Cat. |
| MDRS Y | 0–3 | 2.93 | 3 | Ordered Cat. |

functioning outcomes as well as a concurrent set of brain MRI measurements. We defined concurrent as within six months (before or after) of the neuropsychological testing date. As one can see from the summary statistics, the outcomes vary greatly in their number of categories as well as in their difficulty. For many of the binary outcomes as well as the MDRS outcomes E and V, the mean and median scores are very close to the largest possible score.

To illustrate the challenges of modeling cognitive outcomes from the SIVD study parametrically, we describe two items in more detail. For MDRS outcome E, participants are given one minute and are asked to name as many items found in supermarkets as they can. The participant's score is the number of valid items named, censored at 20. A histogram of observed scores for this outcome in Figure 1(a) shows some evidence of a ceiling effect for this item. Similarly, Figure 1(b) depicts a histogram of observed scores for MDRS outcome W that asks a participant to compare words and identify similarities. Although the description in this case does not suggest right-censoring, there is also some evidence of a ceiling effect in the histogram. We might treat MDRS outcome W as right-censored rather than an ordered categorical outcome in a parametric approach. These are just two examples that illustrate ambiguities in specifying appropriate parametric distributions for each cognitive outcome in the SIVD study. To bypass this specification, yet




(a) MDRS E (b) MDRS W

FIG. 1. *Histograms of scores for MDRS E and W items.*

still model the interdependencies among test items, we used the hierarchical semiparametric latent variable model.

We are interested in modeling the relationship between the primary factor and the volume of white matter hyperintensities located in the frontal lobe of the brain. Controlling for other covariates, we specified the mean of the primary factor as

$$\mathrm{E}[\eta_{i1}] = \beta_1 \mathrm{Sex}_i + \beta_2 \mathrm{Educ}_i + \beta_3 \mathrm{Age}_i + \beta_4 \mathrm{Vol}_i + \beta_5 \mathrm{WMH}_i, \tag{36}$$

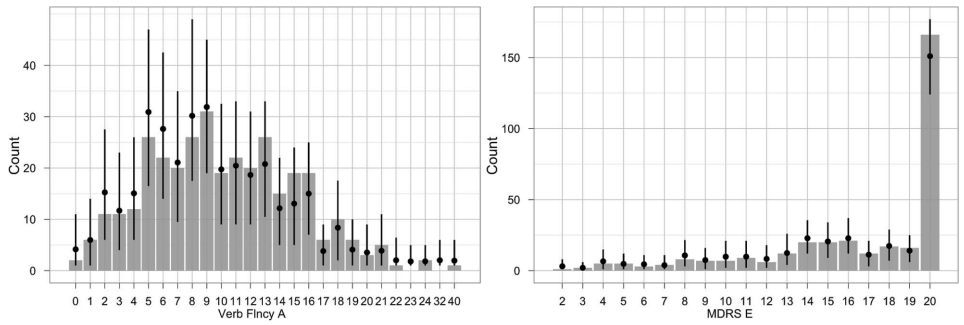
where Sex is the participant's sex (Female = 1, Male = 0), Educ is the number of years of education, Vol is the total brain volume of the participant, and WMH is the frontal white matter hyperintensity volume. We used standardized versions of the continuous predictor variables. Table 2 displays some summary statistics for these covariates by different levels of frontal white matter hyperintensity volume.

*One-factor semiparametric model*. We started our analysis by examining the $Q = 1$ model with a single latent factor explaining interdependencies among the test items. To estimate the model, we utilized the parameter-expanded Gibbs sampling algorithm. Even though we found this approach to be more efficient than the

TABLE 2

*Mean and SD for covariates by level of frontal WMH. The range of frontal WMH measurements was partitioned to obtain three similarly sized groups*

| | **Frontal WMH (cc)** | | |
|---|---|---|---|
| | **0–5** | **>5–11** | **>11** |
| No. participants | 113 | 112 | 116 |
| Age (Yrs) | 68.49 (8.65) | 74.82 (6.72) | 79.44 (6.23) |
| Education (Yrs) | 15.36 (2.95) | 15.12 (3.01) | 14.32 (3.12) |
| Total brain volume (cc) | 1196.2 (115.81) | 1231.65 (114.78) | 1218.34 (138.13) |




(a) Verbal Fluency A (b) MDRS E

FIG. 2. *Histograms of the observed scores for the Verbal Fluency A and MDRS E. The black points indicate the mean count across replicated data sets for each score. The black vertical segment indicates the interval from the* 2.5% *to* 97.5% *quantiles across replicated data sets.*

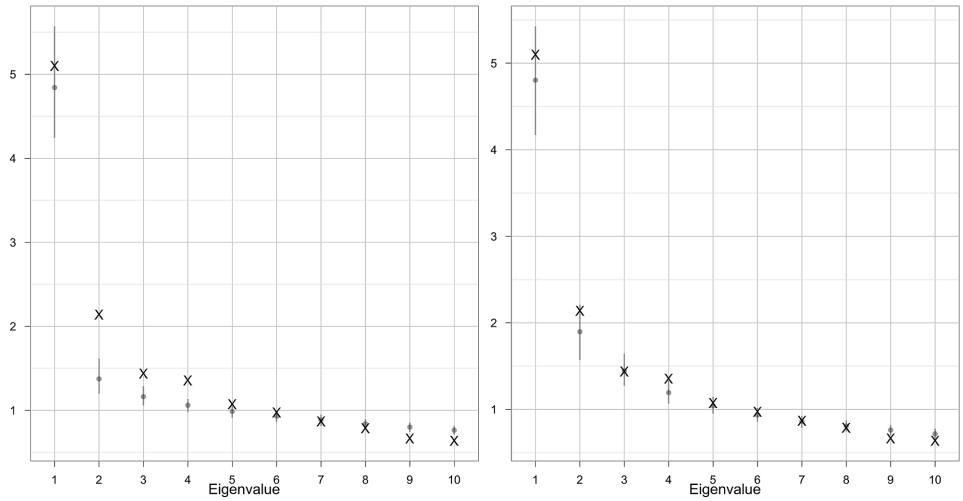
standard Gibbs sampler, we still observed high autocorrelation within the chains for factor loadings. We drew 50,000 MCMC samples and discarded the first half as burn-in. We used trace plots and the Geweke [Geweke (1992)] and Raftery–Lewis [Raftery and Lewis (1995)] diagnostic tests to assess convergence.

Table 4 displays posterior summaries for the regression coefficients, $\boldsymbol{\beta}$. We observed a negative relationship between the primary factor and frontal white matter hyperintensity volume. The accompanying 95% posterior credible interval $(-0.466, -0.205)$ did not contain zero, suggesting a negative association between frontal white matter hyperintensity volume and the primary factor.

We evaluated model fit using posterior predictive model checks. We began by examining the fit of the marginal distributions. Figure 2 displays the histograms of observed responses for Verbal Fluency outcome $A$ and MDRS outcome $E$ along with posterior predictive summaries. In each case, the model appeared to do a satisfactory job of approximating the data. We found similarly good approximations of the marginal distributions in the observed data for the other outcomes as well.

We assessed the model's ability to replicate the observed dependence structure in the data at a global level by examining the eigenvalues of the observed rank correlation matrix [Figure 3(a)]. The eigenvalues of correlation matrices form the basis of heuristic tests in factor analysis such as the latent root criterion [Guttman (1954)] or the scree test [Cattell (1966)] that determine the number of factors to include in the model. The first eigenvalue was well approximated by the model but the subsequent eigenvalues indicated model misfit, suggesting that additional factors may be necessary to more accurately represent the dependence structure in the data.

We reviewed the pairwise rank correlations to better understand the shortcomings of the single factor model and direct the next steps in our model building process. Figures 4(a) and 4(b) display the pairwise correlation plots for the MDRS J and Visual Span Backward outcomes for the single factor model. In both cases,

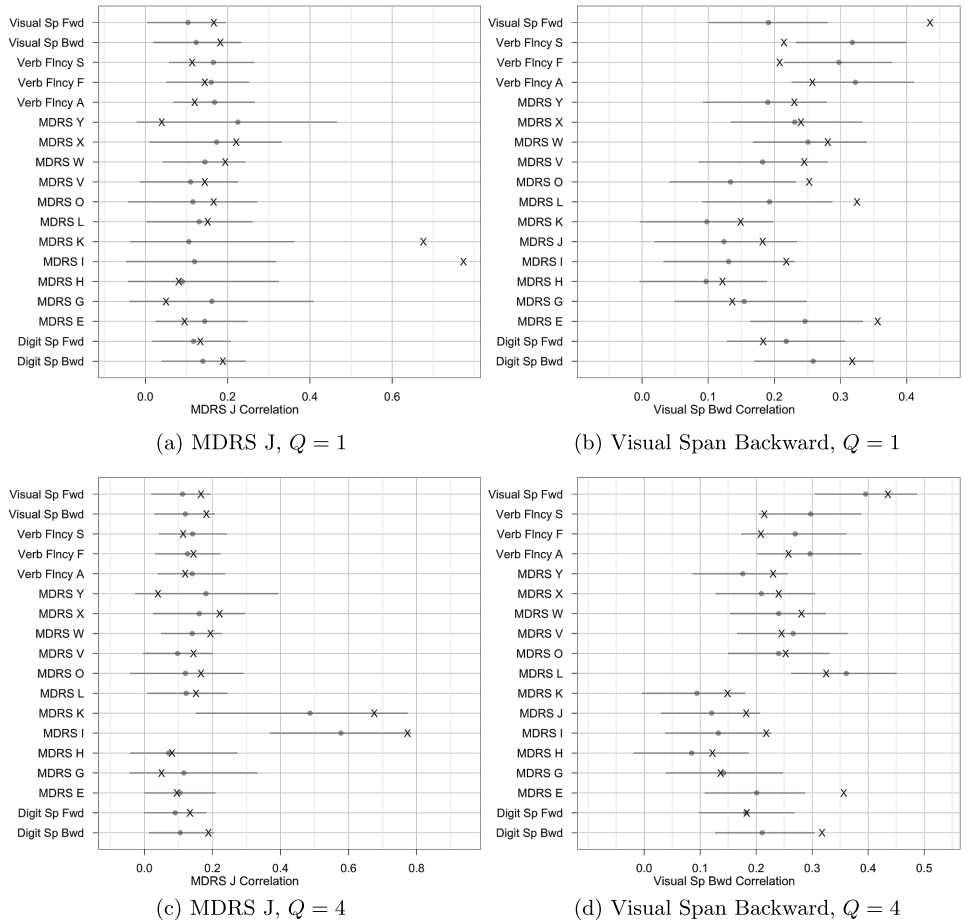



(a) $Q = 1$ (b) $Q = 4$

FIG. 3. *Eigenvalue plots for the* $Q = 1$ *and* $Q = 4$ *models. The mean posterior prediction* (*grey point*) *and* 95% *posterior prediction intervals* (*grey line segment*) *of the top ten eigenvalues calculated using replicated data from the single factor* ($Q = 1$) *and bifactor* ($Q = 4$) *models. Eigenvalues computed from the observed data are denoted by a black "X."*

the model fit the majority of the pairwise correlations well. However, in each case, there were a few outcomes with poorly fitted correlations. For MDRS J, the model did not appear to fully capture the correlation with the conceptually-related MDRS I and K; all three of these outcomes ask participants to repeat alternating movements of some type. Likewise, for Visual Span Backward, the correlation with Visual Span Forward was not accurately approximated by the single factor model. In addition, the correlations between Visual Span Backward and the MDRS outcomes L and O were not well approximated. MDRS outcomes L and O involve copying drawings and, in this sense, also incorporate a visual component that may be the source of the residual correlation between the outcomes. The observation that the lack of fit was present among conceptually related outcomes (e.g., outcomes that are parts of a subtest or a subscale) is consistent with the notion that possible secondary factors may impact item correlations in addition to the general executive functioning factor. Thus, our next step was to consider the class of bifactor models.

*Bifactor semiparametric model*. To choose a secondary factors structure in a bifactor model, we applied an iterative process. During one iteration, we examined all pairwise correlations for the lack of fit, specified secondary factors to account for residual correlation, refit the model and checked the fit of this new model. Ultimately, we specified a bifactor model with one general cognitive ability factor and 3 secondary factors (for a total of $Q = 4$) as listed in Table 3. It is important to note that, although we identified these secondary factors using the posterior




(a) MDRS J, $Q = 1$

(b) Visual Span Backward, $Q = 1$

(c) MDRS J, $Q = 4$

(d) Visual Span Backward, $Q = 4$

FIG. 4. *Pairwise correlation plots for the single factor* ($Q = 1$) *and bifactor models* ($Q = 4$). *Each pairwise correlation plot depicts the mean posterior prediction* (*grey point*) *and* 95% *posterior prediction intervals* (*grey line segment*) *for Kendall's* $\tau$ *values calculated using replicated data. Kendall's* $\tau$ *values computed from the observed data are denoted by a black "X."*

predictive model checks, they nevertheless have substantive interpretations as they link conceptually related outcomes. The second factor loads on MDRS outcomes I, J and K, test items that all involve repetition of alternating movements. The third factor loads on the Visual Span outcomes and MDRS outcomes L, O and V. These test items all include visual or drawing components. The fourth factor links three MDRS outcomes that ask participants to identify similarities and dissimilarities.

For the semiparametric bifactor model with $Q = 4$, we drew 500,000 MCMC samples and discarded the first 50,000 as burn-in. We kept every 50th draw, leaving us with 9000 posterior draws. As with the single factor model, we checked convergence using trace plots and the Geweke [Geweke (1992)] and Raftery–Lewis

TABLE 3
*Proposed factor structure for SIVD executive functioning outcomes. ∗ indicates a nonzero factor loading to be estimated*

| | Outcomes | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Factor** | **MDRS G** | **MDRS H** | **MDRS I** | **MDRS J** | **MDRS K** | **MDRS L** | **MDRS O** | **Digit Sp Fwd** | **Digit Sp Bwd** | **Visual Sp Fwd** | **Visual Sp Bwd** | **Verb Flncy F** | **Verb Flncy A** | **Verb Flncy S** | **MDRS E** | **MDRS V** | **MDRS W** | **MDRS X** | **MDRS Y** |
| 1 | ∗ | ∗ | ∗ | ∗ | ∗ | ∗ | ∗ | ∗ | ∗ | ∗ | ∗ | ∗ | ∗ | ∗ | ∗ | ∗ | ∗ | ∗ | ∗ |
| 2 | 0 | 0 | ∗ | ∗ | ∗ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 0 | ∗ | ∗ | 0 | 0 | ∗ | ∗ | 0 | 0 | 0 | 0 | ∗ | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ∗ | ∗ | ∗ |

[Raftery and Lewis (1995)] diagnostic tests. Convergence was satisfactory but, compared to the single factor model, the mixing was considerably slower for a few of the secondary factors that exhibited high levels of autocorrelation. We should also note that the speed of convergence was influenced by the choice of hyperparameters for $\mathbf{\Sigma}$ and $\mathbf{\Psi}$ in the parameter expanded model.

The bifactor model represented the dependence structure of the observed responses better. Figure 3(b) shows that the bifactor model provided a good fit to the observed eigenvalues well beyond the first eigenvalue. As can be seen in Figures 4(c) and 4(d), the bifactor model did a better job of replicating the pairwise rank correlations compared to the single factor model. In Figure 4(c), one may also notice the larger posterior credible intervals for the pairwise rank correlations of MDRS I with MDRS J and MDRS K. Referring back to Table 1, we see that almost all participants answered these items correctly. As a result, there is less information to estimate the secondary factor that accounts for the residual correlation among these three items and, in the wide credible intervals, we see the subsequent imprecision.

Table 4 displays posterior summaries for the regression parameters. We saw little change in our estimate for the parameter of interest, $\beta_5$, the coefficient for frontal WMH in adding additional factors. Thus, our substantive conclusion regarding the association between an individual's executive functioning and the volume of white matter hyperintensities in the frontal region of the brain remains the same whether we use the one-factor model or the better fitting bifactor model. Based on our semiparametric latent variable model, we expect a 1SD increase in frontal white matter hyperintensity volume to be associated with a 0.335SD decrease in the primary factor. In examining the other coefficients, we see that none of the 95% posterior credible intervals have shifted to the extent that we would

TABLE 4
*Posterior summaries for regression coefficients for single factor, $Q = 1$, and bifactor, $Q = 4$, models*

| | $Q = 1$ | | | $Q = 4$ | | |
|---|---|---|---|---|---|---|
| **Coefficient** | **Mean** | **Median** | **95% CI** | **Mean** | **Median** | **95% CI** |
| Sex | 0.234 | 0.233 | (−0.061, 0.516) | 0.155 | 0.152 | (−0.134, 0.440) |
| Education | 0.354 | 0.354 | (0.232, 0.479) | 0.325 | 0.325 | (0.206, 0.446) |
| Age | −0.126 | −0.126 | (−0.246, 0.004) | −0.078 | −0.078 | (−0.201, 0.044) |
| Total brain Vol. | 0.069 | 0.069 | (−0.080, 0.215) | 0.046 | 0.045 | (−0.096, 0.194) |
| Frontal WMH Vol. | −0.330 | −0.328 | (−0.466, −0.205) | −0.335 | −0.336 | (−0.464, −0.208) |

alter our posterior belief about whether zero is a plausible value for the parameter. However, the coefficients for sex, age and total brain volume did decrease by 30–40% in magnitude.

**6. Discussion.** In this paper we have developed a semiparametric latent variable model for multivariate mixed outcome data. This model, unlike common parametric latent variable modeling approaches for mixed outcome data [Dunson (2003), Moustaki and Knott (2000), Sammel, Ryan and Legler (1997), Shi and Lee (1998)], does not require the specification of conditional distributions for each outcome given the latent variables. When a data set combines a variety of mixed outcomes, picking appropriate conditional distributions for each outcome encountered in real data, extending the parametric models to account for all cases of distributions, and extending estimation methods appropriately can be labor-intensive. Moreover, specification of outcome conditional distributions given the latent variables may be of little interest by itself in any research setting where the main question is in the relationship between a common factor (or factors) and a covariate of interest. Our proposed semiparametric latent variable framework allows one to model interdependencies among observed mixed outcome variables by specifying an appropriate latent variable model while, at the same time, avoiding the specification of outcome distributions conditional on the common latent variables. We have demonstrated this approach for the single-factor and bifactor models, incorporating a covariate effect on the general factor.

The extended rank likelihood can readily be employed with other latent variable models, including item response theory models [van der Linden and Hambleton (1997)] and structural equation models [Bollen (1989)]. In structural equation models, the focus is often on characterizing the relationship between latent variables and/or between latent variables and fixed covariates as in the case of our hierarchical model. In such cases where the focus is not on the loadings or outcome-related parameters, the proposed semiparametric approach would be quite useful in dealing with mixed outcome data. However, the extended rank likelihood may not be as useful in cases where outcome-specific parameters on the scale of the

observed outcomes are of interest. In item response theory models, one is often interested in examining the item difficulty and discrimination parameters to better understand the characteristics of individual test questions. The difficulty parameter, the analogue to the specific mean in the factor model, is not directly identifiable with the extended rank likelihood approach. Nonetheless, one could still carry out posterior inference by relying on the relationship between the difficulty parameter and the latent trait. For example, in a two-parameter item response theory model for binary outcomes, the probability of a positive response when the factor score is set to zero is a one-to-one function of the difficulty parameter. Such an alternative, however, may render the semiparametric approach less convenient for a practitioner who is primarily interested in parameters characterizing the properties of individual outcomes.

We employed the semiparametric latent variable model to study the association between the volume of white matter hyperintensities in the frontal lobe and cognitive testing outcomes related to executive functioning from the Subcortical Ischemic Vascular Dementia (SIVD) study. The semiparametric latent variable model allowed us to analyze the mixed cognitive testing outcomes without requiring the specification of parametric distributions for the outcomes conditional on the latent variables. It has been hypothesized that a greater volume of frontal lobe white matter hyperintensities will be associated with worse executive functioning. Consistent with this hypothesis, we found a negative association between the primary factor in our model and the volume of white matter hyperintensities.

Our model selection process was guided by substantive beliefs that associations among items in the cognitive testing data are primarily driven by the main latent factor but can potentially be influenced by secondary latent factors due to local dependencies among groups of related items. Thus, we started our model-building process by fitting the one-factor semiparametric model and relied on posterior predictive model checks to evaluate model misfit and to guide us in identifying a secondary factor structure for the bifactor model. Our posterior predictive checks approach can therefore be thought of as a method of exploratory bifactor analysis when the secondary factor structure is not known in advance [Jennrich and Bentler (2011)]. It also provides a mechanism by which statistical methodologists can work together with substantive experts to develop models that are theoretically justified and that are consistent with the data. We note, however, that the main conclusion about the association between executive functioning and regional brain volumes was not affected much by the choice of a better fitting bifactor model over the single factor model in our case.

While our proposed model selection process is somewhat ad-hoc, one could explore the use of more formal model fit criteria, other model selection methods or a fully Bayesian approach to determine the factor structure for our semiparametric model. For example, one could use the methods of Knowles and Ghahramani (2011) and Rai and Daumé III (2009) to incorporate the Indian Buffet Process

prior to simultaneously estimate the loadings, the loadings structure and the number of factors. Within the bifactor model framework, Jennrich and Bentler (2011) recently proposed using a rotation criterion to explore the secondary factor structure. Dunson et al. (2006) presented a Bayesian model averaging approach that accounts for the uncertainty in the number of factors.

Overall, in our work with the cognitive testing data, we found that the semiparametric model was more elegant and much easier in implementation than the standard parametric approaches for mixed outcome data.

**Acknowledgments.** The authors would like to thank Peter Hoff, S. McKay Curtis, Thomas Richardson, Dan Mungas, Laura Gibbons and the Cognitive Outcomes with Advanced Pyschometrics group, University of Washington, for many helpful discussions and comments on earlier versions of this work. In addition, the reviewers of the original submission provided valuable feedback that have strengthened the final version.

## REFERENCES

ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley, Hoboken, NJ. MR1990662

BARTHOLOMEW, D., KNOTT, M. and MOUSTAKI, I. (2011). *Latent Variable Models and Factor Analysis*: *A Unified Approach*, 3rd ed. Wiley, Chichester. MR2849614

BOLLEN, K. A. (1989). *Structural Equations with Latent Variables*. Wiley, New York. MR0996025

CARDENAS, V. A., EZEKIEL, F., DI SCLAFANI, V., GOMBERG, B. and FEIN, G. (2001). Reliability of tissue volumes and their spatial distribution for segmented magnetic resonance images. *Psychiatry Research*: *Neuroimaging* **106** 193–205.

CATTELL, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research* **1** 245–276.

CHUI, H. C. (2007). Subcortical ischemic vascular dementia. *Neurol. Clin.* **25** 717–740, vi.

CHUI, H. C., ZAROW, C., MACK, W. J., ELLIS, W. G., ZHENG, L., JAGUST, W. J., MUNGAS, D., REED, B. R., KRAMER, J. H., DECARLI, C. C. et al. (2006). Cognitive impact of subcortical vascular and Alzheimer's disease pathology. *Annals of Neurology* **60** 677.

CONGDON, P. (2003). *Applied Bayesian Modelling*. Wiley, Chichester. MR1990543

CONGDON, P. (2006). *Bayesian Statistical Modelling*, 2nd ed. Wiley, Chichester. MR2281386

DOBRA, A. and LENKOSKI, A. (2011). Copula Gaussian graphical models and their application to modeling functional disability data. *Ann. Appl. Stat.* **5** 969–993. MR2840183

DUNN, J. E. (1973). A note on a sufficiency condition for uniqueness of restricted factor matrix. *Psychometrika* **38** 141–143. MR0345326

DUNSON, D. B. (2003). Dynamic latent trait models for multidimensional longitudinal data. *J. Amer. Statist. Assoc.* **98** 555–563. MR2011671

DUNSON, D. B. et al. (2006). Efficient Bayesian model averaging in factor analysis. Technical report, Duke Univ., Durham, NC.

EROSHEVA, E. and CURTIS, S. M. (2011). Specification of rotational constraints in Bayesian confirmatory factor analysis. Technical Report No. 589, Univ. Washington, Seattle, WA.

GEWEKE, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics*, 4 (*PeñíScola*, 1991) (J. M. Bernardo, J. Berger, A. P. Dawid and J. F. M. Smith, eds.) 169–193. Oxford Univ. Press, New York. MR1380276

GEWEKE, J. and ZHOU, G. (1996). Measuring the pricing error of the arbitrage pricing theory. *Review of Financial Studies* **9** 557–587.

GHOSH, J. and DUNSON, D. B. (2008). Bayesian model selection in factor analytic models. In *Random effect and latent variable model selection* 151–163. Springer, New York.

GHOSH, J. and DUNSON, D. B. (2009). Default prior distributions and efficient posterior computation in Bayesian factor analysis. *J. Comput. Graph. Statist.* **18** 306–320. MR2749834

GRUHL, J., EROSHEVA, E. and CRANE, P. (2010). Analyzing cognitive testing data with extensions of item response theory models. Presented at the Joint Statistical Meetings, Vancouver, Canada, August 3, 2010.

GRUHL, J., EROSHEVA, E. and CRANE, P. (2011). A semiparametric Bayesian latent trait model for multivariate mixed type data. In *International Meeting of the Pyschometric Society*.

GUTTMAN, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika* **19** 149–161. MR0091235

HACHINSKI, V., IADECOLA, C., PETERSEN, R. C., BRETELER, M. M., NYENHUIS, D. L., BLACK, S. E., POWERS, W. J., DECARLI, C., MERINO, J. G., KALARIA, R. N. et al. (2006). National institute of neurological disorders and stroke—Canadian stroke network vascular cognitive impairment harmonization standards. *Stroke* **37** 2220–2241.

HOFF, P. D. (2007). Extending the rank likelihood for semiparametric copula estimation. *Ann. Appl. Stat.* **1** 265–283. MR2393851

HOFF, P. D. (2009). *A First Course in Bayesian Statistical Methods*. Springer, New York. MR2648134

HOLZINGER, K. J. and SWINEFORD, F. (1937). The bi-factor method. *Psychometrika* **2** 41–54.

JENNRICH, R. I. (1978). Rotational equivalence of factor loading matrices with specified values. *Psychometrika* **43** 421–426. MR0514727

JENNRICH, R. I. and BENTLER, P. M. (2011). Exploratory bi-factor analysis. *Psychometrika* **76** 537–549. MR2851500

JÖRESKOG, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika* **34** 183–202.

KLÜPPELBERG, C. and KUHN, G. (2009). Copula structure analysis. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71** 737–753. MR2749917

KNOWLES, D. and GHAHRAMANI, Z. (2011). Nonparametric Bayesian sparse factor models with application to gene expression modeling. *Ann. Appl. Stat.* **5** 1534–1552. MR2849785

KUCZYNSKI, B., TARGAN, E., MADISON, C., WEINER, M., ZHANG, Y., REED, B., CHUI, H. C. and JAGUST, W. (2010). White matter integrity and cortical metabolic associations in aging and dementia. *Alzheimer's and Dementia* **6** 54–62.

LIU, C., RUBIN, D. B. and WU, Y. N. (1998). Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika* **85** 755–770. MR1666758

LIU, J. S. and WU, Y. N. (1999). Parameter expansion for data augmentation. *J. Amer. Statist. Assoc.* **94** 1264–1274. MR1731488

LOKEN, E. (2005). Identification constraints and inference in factor models. *Struct. Equ. Model.* **12** 232–244. MR2135915

LOPES, H. F. and WEST, M. (2004). Bayesian model assessment in factor analysis. *Statist. Sinica* **14** 41–67. MR2036762

MILLSAP, R. E. (2001). When trivial constraints are not trivial: The choice of uniqueness constraints in confirmatory factor analysis. *Struct. Equ. Model.* **8** 1–17.

MORRIS, J. C. (1993). The Clinical Dementia Rating (CDR): Current version and scoring rules. *Neurology* **43** 2412–2414.

MORRIS, J. C. (1997). Clinical dementia rating: A reliable and valid diagnostic and staging measure for dementia of the Alzheimer type. *Int. Psychogeriatr.* **9 Suppl 1** 173–176; discussion 177–178.

MOUSTAKI, I. and KNOTT, M. (2000). Generalized latent trait models. *Psychometrika* **65** 391–411. MR1792703

MUNGAS, D., HARVEY, D., REED, B. R., JAGUST, W. J., DECARLI, C., BECKETT, L., MACK, W. J., KRAMER, J. H., WEINER, M. W., SCHUFF, N. et al. (2005). Longitudinal volumetric MRI change and rate of cognitive decline. *Neurology* **65** 565–571.

MURAKI, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Appl. Psychol. Meas.* **16** 159.

MURRAY, J. S., DUNSON, D. B., CARIN, L. and LUCAS, J. E. (2013). Bayesian Gaussian copula factor models for mixed data. *J. Amer. Statist. Assoc.* **108** 656–665.

PETTITT, A. N. (1982). Inference for the linear model using a likelihood based on ranks. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **44** 234–243. MR0676214

RAFTERY, A. E. and LEWIS, S. M. (1995). The number of iterations, convergence diagnostics and generic Metropolis algorithms. In *Practical Markov Chain Monte Carlo* (W. R. Gilks, D. J. Spiegelhalter and S. Richardson, eds.). Chapman & Hall, London, UK.

RAI, P. and DAUMÉ III, H. (2009). The infinite hierarchical factor regression model. Available at arXiv:0908.0570.

REISE, S. P., MORIZOT, J. and HAYS, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Qual. Life Res.* **16 Suppl 1** 19–31.

SAMEJIMA, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement* **34** 1–100.

SAMMEL, M. D., RYAN, L. M. and LEGLER, J. M. (1997). Latent variable models for mixed discrete and continuous outcomes. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **59** 667–678.

SHI, J. Q. and LEE, S. Y. (1998). Bayesian sampling-based approach for factor analysis models with continuous and polytomous data. *British J. Math. Statist. Psych.* **51** 233–252.

SKRONDAL, A. and RABE-HESKETH, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman & Hall/CRC, Boca Raton, FL. MR2059021

STEPHENS, M. (2000). Dealing with label switching in mixture models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **62** 795–809. MR1796293

VAN DER LINDEN, W. J. and HAMBLETON, R. K., eds. (1997). *Handbook of Modern Item Response Theory*. Springer, New York. MR1601043

WEST, M. (1987). On scale mixtures of normal distributions. *Biometrika* **74** 646–648. MR0909372

J. GRUHL
E. A. EROSHEVA
DEPARTMENT OF STATISTICS
UNIVERSITY OF WASHINGTON
SEATTLE, WASHINGTON 98195-4322
USA
E-MAIL: gruhlj@stat.washington.edu
elena@stat.washington.edu

P. K. CRANE
DEPARTMENT OF MEDICINE
HARBORVIEW MEDICAL CENTER
325 NINTH AVENUE
CAMPUS BOX 359780
SEATTLE, WA 98104
USA
E-MAIL: pcrane@u.washington.edu