Joint Bayesian Modeling of Binomial and Rank Data for Primate Cognition

Abstract

In recent years, substantial effort has been devoted to methods for analyzing data containing mixed response types, but such techniques typically do not include rank data among the response types. Some unique challenges exist in analyzing rank data, particularly when ties are prevalent. We present techniques for jointly modeling binomial and rank data using Bayesian latent variable models. We apply these techniques to compare the cognitive abilities of non-human primates based on their performance on 17 cognitive tasks scored on either a rank or binomial scale. In order to jointly model the rank and binomial responses, we assume that responses are implicitly determined by latent cognitive abilities. We then model the latent variables using random effects models, with identifying restrictions chosen to promote parsimonious prior specification and model inferences. Results from the primate cognitive data are presented to illustrate the methodology. Our results suggest that the ordering of the cognitive abilities of species varies significantly across tasks, suggesting a partially independent evolution of cognitive abilities in primates.

Key phrases: mixed response; latent performance; interanimal variability.

1 Introduction

The motivation for the methodology presented in this paper arises from a study of the cognitive abilities of non-human primates in a variety of cognitive tasks, or assessments. Test data were collected from a number of research groups on 99 primates from seven distinct species: chimpanzees (*Pan troglodytes*), bonobos (*Pan paniscus*), gorillas (*Gorilla gorilla*), orangutans (*Pongo abelii*), spider monkeys (*Ateles geoffroyi*), brown capuchin monkeys (*Sapajus apella*), and long-tailed macaques (*Macaca fascicularis*). Seventeen types of assessment were studied, and according to existing literature each assessment was categorized as measuring one of four intelligence domains: inhibition (I), memory (M), transposition (T), and support (S). Most assessments were summarized as the percentage of correct trials and were therefore considered to represent a binomial response. However, two assessments were summarized as the ratio of the proportion of correct responses under experimental to control conditions, while a third assessment measured how long an individual was able to delay a particular response. Because the distribution of responses for these three assessments were highly irregular, a rank trans-

Table 1: The 17 assessments used for studying primate intelligence. Detailed information on all assessments is available in Amici et al. (2012)

No.	Assessment Description	Modeled	Domain
1	A not B (1 trials exp., 1 control)	Rank	Inhibition
2	Middle Cup (2 trials exp., control)	Rank	Inhibition
3	Plexiglas (2 trials)	Binomial	Inhibition
4	Delay Gratification	Rank	Inhibition
5	Swing Door (10 trials)	Binomial	Inhibition
6	Memory Task 30 Seconds (3 trials)	Binomial	Memory
7	Memory Task 30 Minutes (3 trials)	Binomial	Memory
8	Transposition Single Condition (2 trials)	Binomial	Transposition
9	Transposition Double Condition (2 trials)	Binomial	Transposition
10	Transposition Reversed Condition (1 trial)	Binomial	Transposition
11	Transposition Unbaited Condition (1 trial)	Binomial	Transposition
12	Support Cloth Side (6 trials)	Binomial	Support
13	Support Cloth Ripped (6 trials)	Binomial	Support
14	Support Cloth Bridge (6 trials)	Binomial	Support
15	Support Wool Broken (6 trials)	Binomial	Support
16	Support Wool Onto (6 trials)	Binomial	Support
17	Support Wool Touch (6 trials)	Binomial	Support

formation of the original responses were applied to data from these tests. Thus, each of the 17 assessments was considered to have either a binomial or a rank response. Table 1 contains the classifications used for each assessment's response type and domain. We note that none of the animals were tested in all 17 assessments.

We had two primary goals in analyzing these data. First, we were interested in determining whether and to what extent species-level and individual-level effects might explain performance across assessments. Answering this question is important for understanding whether differences in cognitive performance are better explained in terms of inter-individual differences, or whether different primate species have different cognitive abilities. Second, we wanted to know if these effects were broadly applicable across all assessments or whether the effects were domain-specific. Answering this question provides insight into whether the primate mind consists of specialized and independently evolving cognitive modules, or whether a a general factor accounts for the variance in performance across the different cognitive domains. Evidence of domain-specific factors would provide support for the idea that specific cognitive skills evolve independently, possibly in response to specific socio-ecological demands (see Amici et al. 2012; Banerjee et al. 2009).

Full details of the data collection process, historical context, and a summary of results are contained in Amici et al. (2012). We restrict attention herein to those issues most pertinent towards the statistical methodology used to analyze these data. That is, we focus here on techniques for the joint analysis of rank and binomial response data. This framework may be broadly applied by recognizing that any outcome may be transformed to rank data. In particular, it may be advantageous for continuous outcomes with severe skewness or outliers to be converted to rank data, while binomial outcomes may be left unchanged so as to not lose information.

Our methodology relies on data augmentation (Tanner and Wong, 1987) and the introduction of latent variables to represent each individual's (possibly domain-specific) cognitive abililty. Our latent variable formulation is similar to that proposed for binomial and ordinal data by Albert and Chib (1993), and for multi-rater rank data by Johnson et al. (2002). In the context of a binary response, for example, an observed value y_i may be viewed as an indicator of whether or not a corresponding latent variable z_i exceeds a threshold parameter. Albert and Chib (1993) demonstrated that the z_i's themselves can be readily modeled in a Bayesian framework to allow dependence on covariates.

While much attention has been given to Bayesian latent variable models for Bernoulli responses, less attention has been devoted to corresponding models for rank data. Hoff (2007) used ranks to estimate copula, thereby allowing the marginal distributions in multivariate data to be unspecified while still modeling dependence. Murray et al. (2013) built on Hoff (2007) by forming a Gaussian copula factor model to jointly handle rank response and other response types. While these approaches permit ties by considering the data to be only partially ordered, they do not provide a model for the probability that two outcomes will be tied. Johnson et al. (2002) introduced a model for primate cognition based only on rank data, and they explicitly modeled the probability that any two given observations were tied. The key features of this strategy are (1) that the ordering of the latent variables must be consistent with the ordering of the observed outcomes, and (2) that the probability two individuals k and k' have a tied response is determined by the proximity of the latent variables:

$$Pr(y_k = y_{k'} | z_k, z_{k'}, \kappa) = \exp(-|z_k - z_{k'}|/\kappa).$$

Dunson (2000), Gueorguieva and Agresti (2001), and Gruhl et al. (2013), among others, have used latent variable models to model multivariate observations with mixed response types. Many authors have jointly modeled ordinal and continuous data in a Bayesian framework with data augmentation (e.g., Quinn 2004; Fahrmeir and Raach 2007). Existing techniques, however, do not deal explicitly with modeling rank data in conjunction with other types of responses while modeling the probability of ties in the rank data. In the next section, we show that there is a natural data augmentation approach that allows us to jointly model rank and binomial data with random effects models. This approach facilitates prior specification and model interpretation because of our choice for the identifying constraint. Furthermore, our approach presents no excessive computational burden when tied ranks are prevalent—a recognized hurdle (Johnson et al. 2002) faced by some other approaches to rank data.

2 Methodology

We build upon the previous latent variable models to specify our joint distribution on rank and binomial data. Sections 2.1–2.3 describe the likelihood function and prior distributions proposed in our model framework. Section 2.4 briefly outlines a Metropolis-within-Gibbs MCMC algorithm that may be used for inference on the posterior distribution; further details on the algorithm appear in the supplementary materials.

2.1 Model Specification

Suppose that data are collected from I subjects. Suppose also that data are available from J distinct assessments, with each assessment having either a binomial outcome or a rank outcome. If assessment j has a binomial outcome, let T_j denote the number of trials for the outcome; if assessment j has a rank outcome define $T_j \equiv 1$. Then the response data can be characterized as a collection of y_{ijt} values with $i = 1, \ldots, I$ for the subject number, $j = 1, \ldots, J$ for the assessment number, and $t = 1, \ldots, T_j$ for the trial number. When referring to rank assessments, the t subscript may be omitted (i.e., y_{ij} instead of y_{ijt}). Rank data are assumed to be ordered so that higher values of y_{ij} reflect better performance. Similarly, a binary outcome of 1 represents higher functioning.

In order to distinguish assessments involving ranks from those involving binomial outcomes, let $B(j) = 1$ if assessment j has a binomial response and $B(j) = 0$ if it has a rank response. Furthermore, for each assessment j such that $B(j) = 0$, let $C(j)$ be the number of animals that participated in that assessment. We assume that the selection of animals for participation in each experiment was based on logistical considerations (e.g., which animals were available for and cooperative in testing) and was unrelated to the difficulty and nature of the assessment mechanism.

2.2 Likelihood

For each observed outcome y_{ijt} we associate a latent variable z_{ijt}. Given these latent variables and all other model parameters, we assume that all observations are conditionally independent. For each Bernoulli outcome, we assume that $y_{ijt} = 1$ if and only if $z_{ijt} > \tau_j$, where τ_j represents a discrimination parameter for assessment j. For rank data, we assume that $y_{ij} > y_{kj}$ only if $z_{ij} > z_{kj}$.

To handle ties in rank data, let $y_{(i),j}$ denote the ordered rank statistics for experiment j. Then we define $p_{(ij)}(\kappa_j)$ to be the probability of observing the tie status of the i and $(i+1)$'st order statistics. That is,

$$p_{(ij)}(\kappa_j) = \begin{cases} \exp(-(z_{(i+1),j} - z_{(i),j})/\kappa_j) & \text{if } y_{(i+1),j} = y_{(i),j} \\ 1 - \exp(-(z_{(i+1),j} - z_{(i),j})/\kappa_j) & \text{if } y_{(i+1),j} > y_{(i),j}. \end{cases} \tag{1}$$

Note that Equation (1) differs from the formulation in Johnson et al. (2002) in that it uses an assessment-specific value of κ and therefore is appropriate when the proportion of ties is different across assessments. The incorporation of assessment-specific values of κ also has important implications for modeling rank data when there are large differences in the number of items ranked under different assessments, and in particular when large numbers of items are ranked simultaneously. In such cases, small values of κ account for situations in which few ties are recorded, whereas more moderate values of κ reflect the case in which the central portion of the distribution of ranked items are difficult to distinguish, and many more mid-range items are assigned tied values than are the extreme items. This pattern of ties is consistent with a latent distribution of ties that is unimodal, whereas a disproportionate number of ties in the extremes of the ranked values suggests a multimodal distribution of latent traits or some other model deficiency.

Finally, let $w_{ijt} = 1$ if Y_{ijt} is observed and 0 otherwise, with the convention that $0^0 = 1$.

With this notation and modeling assumptions, the likelihood function for the combined rank and binomial data can be written as

$$
\begin{aligned}
f(\mathbf{y}|\mathbf{z},\boldsymbol{\tau},\boldsymbol{\kappa}) = & \left[\prod_{j:B(j)=1} \prod_{i=1}^{I} \prod_{t=1}^{T_j} \left(1(\{y_{ijt} = 0 \cap z_{ijt} \leq \tau_j\} \cup \{y_{ijt} = 1 \cap z_{ijt} > \tau_j\})\right)^{w_{ijt}} \right] \\
& \times \left[\prod_{j:B(j)=0} \left(\prod_{i=1}^{C(j)-1} p_{(ij)}(\kappa_j) \right) \left(\prod_{i=1}^{I} \prod_{i':z_{i'j}<z_{ij}} \left(1(y_{i'j} \leq y_{ij})\right)^{w_{ij}w_{i'j}} \right) \right]
\end{aligned} \tag{2}
$$

The first part of the likelihood is for binomial data (Albert and Chib 1993) and the second is for rank data (see Johnson et al. 2002). This novel combination of likelihoods for these two response types is significant because it means both rank response and binomial response data depend on latent variables that can be modeled in a common framework.

2.3 Hierarchical Model Priors

In Section 2.2 we specified the likelihood distribution and modeled its dependence on several parameters and latent variables. In order to complete the model framework, we now describe the prior distribution of the latent variables and parameters, focusing first on the latent variables.

As others have noted for rank data and binomial data separately, the latent variables $\mathbf{z}$ must have their location and scale established through informative priors or constraints because the model likelihood is invariant to location and scale transformations of these unobserved quantities. Various possibilities exist to ensure identifiability. Among these options, we feel that it is preferable to specify the model so that all latent variables have the same marginal mean and variance under the prior distribution. This facilitates the interpretation and comparison of model parameters.

To illustrate our proposed scaling, suppose there are P random effects, and the p^{th} random effect has L_p levels. Let $u_{p,l(p,i,j)}$ represent the p^{th} random effect at level l, where the level l can depend on the random effect, subject, assessment, or some combination thereof. With this notation, we assume that

$$z_{ijt} = \sum_{p=1}^{P} u_{p,l(p,i,j)} + \epsilon_{ijt} \tag{3}$$

We further assume that $\boldsymbol{\epsilon}|\sigma_\epsilon^2 \sim N(\mathbf{0}, \sigma_\epsilon^2\mathbf{I})$, that $\mathbf{u}_p|\sigma_{u,p}^2 \sim N(\mathbf{0}, \sigma_{u,p}^2\mathbf{I})$ for $p = 1, 2, \ldots, P$, and that all random effects and error terms are conditionally independent. It follows that

$$z_{ijt}|\{\sigma_\epsilon^2, \sigma_{u,1}^2, \ldots, \sigma_{u,P}^2\} \sim N(0, \sigma_\epsilon^2 + \sum_p \sigma_{u,p}^2),$$

which means that a constraint of the form $\sigma_\epsilon^2 + \sum_p \sigma_{u,p}^2 = 1$ ensures that all latent variables have the same marginal expectation and variance, regardless of P. With this constraint, each random effect's variance parameter can be interpreted as the proportion of total variability in $\mathbf{z}$ that is attributed to that random effect. This constraint also simplifies the prior specification of hyperparameters for variance parameters across models, as we demonstrate shortly.

Using the generalized form for the random effects model, the joint prior density is denoted by

$$\pi(\mathbf{z}, \mathbf{u}_1, \ldots, \mathbf{u}_P, \sigma_{u,1}^2, \ldots, \sigma_{u,P}^2, \sigma_\epsilon^2, \boldsymbol{\tau}, \boldsymbol{\kappa}). \tag{4}$$

We assume that this prior density can be factored according to

$$\pi(\mathbf{z}|\mathbf{u}_1, \ldots, \mathbf{u}_P, \sigma_\epsilon^2)\pi(\mathbf{u}_1|\sigma_{u,1}^2) \cdots \pi(\mathbf{u}_P|\sigma_{u,P}^2)\pi(\sigma_{u,1}^2, \ldots, \sigma_{u,P}^2, \sigma_\epsilon^2)\pi(\boldsymbol{\tau})\pi(\boldsymbol{\kappa}),$$

where

$$\pi(\mathbf{z}|\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_P, \sigma_\epsilon^2) = \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{t=1}^{T_j} (2\pi\sigma_\epsilon^2)^{-1/2} \exp(-(z_{ijt} - \sum_p u_{p,l(p,i,j)})^2 / 2\sigma_\epsilon^2) \tag{5}$$

and

$$\pi(\mathbf{u}_p|\sigma_{u,p}^2) = \prod_{l=1}^{L_p} (2\pi\sigma_{u,p}^2)^{-1/2} \exp\left(-u_{1,p}^2/2\sigma_{u,p}^2\right), \quad \text{for} \quad p = 1, \ldots, P. \tag{6}$$

Next, we model the prior distribution of the variance parameters using a Dirichlet density. This enforces the constraint that the variances sum to 1. Let $\boldsymbol{\sigma} \equiv (\sigma_\epsilon^2, \sigma_{u,1}^2, \ldots, \sigma_{u,P}^2)$ and let $\boldsymbol{\alpha}_{\boldsymbol{\sigma}} \equiv (\alpha_\epsilon, \alpha_{u,1}, \ldots, \alpha_{u,P})$. The prior density of $\boldsymbol{\sigma}$ is 0 unless each of the variances is nonnegative and the variances sum to 1; subject to these conditions, the density is

$$\pi(\boldsymbol{\sigma}) = \frac{\Gamma(\alpha_\epsilon + \alpha_{u,1} + \cdots + \alpha_{u,P})}{\Gamma(\alpha_\epsilon)\Gamma(\alpha_{u,1}) \cdots \Gamma(\alpha_{u,P})} (\sigma_\epsilon^2)^{\alpha_\epsilon - 1} (\sigma_{u,1}^2)^{\alpha_{u,1} - 1} \cdots (\sigma_{u,P}^2)^{\alpha_{u,P} - 1}. \tag{7}$$

We complete the prior specification on model parameters by assuming that the components of $\boldsymbol{\kappa} \equiv (\kappa_1, \ldots, \kappa_J)'$ are, *a priori*, conditionally independently drawn from $Gamma(a_j, b_j)$ distributions, and that the components of $\boldsymbol{\tau} \equiv (\tau_1, \ldots, \tau_J)$ are similarly drawn from $Cauchy(m_j, s_j^2)$ distributions. Criteria for fixing the values of the constants a_j, b_j, m_j and s_j depend on substantive considerations. In practice, it might be preferable to select values that are common across assessments because it is not known *a priori* which assessments are most likely to produce ties nor which assessments are the most difficult. Indeed, this is the approach that we took with the primate cognition application presented in Section 3.

2.4 Posterior Inference Using Markov Chain Monte Carlo

It follows that the posterior distribution (up to a normalizing constant) for the combined rank and binomial data can be expressed as

$$f(\mathbf{y}|\mathbf{z}, \boldsymbol{\kappa}, \boldsymbol{\tau})\pi(\mathbf{z}|\mathbf{u}_1, \ldots, \mathbf{u}_P, \sigma_\epsilon^2)\pi(\sigma_\epsilon^2, \sigma_{u,1}^2, \ldots, \sigma_{u,P}^2)\pi(\boldsymbol{\kappa})\pi(\boldsymbol{\tau})\prod_{p=1}^{P}\pi(\mathbf{u}_p|\sigma_{u,p}^2) \tag{8}$$

This model is too complex to study analytically, and so we propose a Metropolis-within-Gibbs sampling technique to simulate posterior samples from it. With one exception, the algorithm updates parameters and latent variables by sampling from full conditional distributions exactly or with a Metropolis-Hastings update using standard methods as described in Johnson et al. (2002) and Albert and Chib (1993). The single innovation of our algorithm lies in the way we update the latent variables for rank data; we introduce an optional location shift of all latent variables in a rank assessment that share the same observed response value. An overview of the MCMC algorithm appears in Table 2; technical details appear in the supplementary materials.

Table 2: Overview of the Metropolis-within-Gibbs sampling algorithm

Step	Action
0	Initialize all unknown values (latent variables, parameters) in support
1	Update $\boldsymbol{\tau}$, $\mathbf{z}$ using modified Cowles' (1996) algorithm and Metropolis-Hastings
2	Update $\boldsymbol{\kappa}$ using Metropolis-Hastings
3	Update $\boldsymbol{\sigma}$ using Metropolis-Hastings
4	Update random effects using complete conditionals
5	Return to Step 1 until sufficiently large number of iterations drawn

3 Application to Primate Cognition

This section applies the previously described framework to data on primate cognition. As described in Section 3.1, we began by fitting a model with four random effects. Based on these results, we also considered several simpler models. The models were compared using metrics given in Section 3.2 that quantify the agreement between the posterior predictive distributions and the observed data. Some results for two models are presented in Section 4. Detailed results for the preferred model are presented in Section 4.2.

3.1 Models for the Primate Data

The primate cognition data served as a motivating application for developing the joint model that allows for random effects. Recall that 99 individuals of 7 primate species were assessed. Each individual was assessed in at least 1 of 17 assessments representing four domains. The combined primate cognitive data set is remarkable because of its large number of observations on the individual level and its inclusion of great apes, New World and Old World monkeys. The data set thus permits investigation of several questions.

- Is there substantial interindividual variability in intelligence after accounting for species effects? If so, how does interindividual variability compare to interspecies variability (on a latent scale)?
- Do assessments from different domains yield different conclusions in comparing cognitive performance between or within species?

To answer these questions, we fit a model with four different random effects: species, species*domain, individual, and individual*domain effects. Rather than the generic notation of Section 2 for random effects (i.e., $u_{p,l(p,i,j)}, p = 1, \ldots, P$), the random effects considered are denoted in a more easily distinguished fashion. Because the species is an individual characteristic, $s(i) = 1, \ldots, 7$ can be used to represent the species of individual i and $u_{s(i)}$ can be used to represent the species effect. The individual effect for individual i is represented by u_i. Each assessment belongs to one of four domains, so $d(j) = 1, \ldots, 4$ is used to represent the domain for assessment j. The species*domain interaction effect is represented by $u_{s(i),d(j)}$. The individual*domain interaction effect is represented by $u_{i,d(j)}$.

Based on the full model results, we considered omitting random effects. Omission of random effects has no impact on the form of the likelihood presented in Section 2.1.

The general prior structure was discussed in detail in Section 2. For the most complex model, M_{SDID}, this becomes

$$\begin{aligned}
&\pi(\mathbf{z}, \boldsymbol{u_s}, \boldsymbol{u_i}, \boldsymbol{u_{s,d}}, \boldsymbol{u_{i,d}}, \boldsymbol{\sigma}, \boldsymbol{\kappa}, \boldsymbol{\tau}) \\
&= \pi(\boldsymbol{\tau})\pi(\boldsymbol{\kappa})\pi(\mathbf{z}|\boldsymbol{u_s}, \boldsymbol{u_i}, \boldsymbol{u_{s,d}}, \boldsymbol{u_{i,d}}, \boldsymbol{\sigma})\pi(\boldsymbol{\sigma}) \\
&\times \pi(\boldsymbol{u_s}|\boldsymbol{\sigma})\pi(\boldsymbol{u_i}|\boldsymbol{\sigma})\pi(\boldsymbol{u_{s,d}}|\boldsymbol{\sigma})\pi(\boldsymbol{u_{i,d}}|\boldsymbol{\sigma})
\end{aligned} \tag{9}$$

This prior specification can be applied to simpler versions of the model if the component prior densities are appropriately restricted.

The vector $\boldsymbol{\sigma}$ contains the variance parameter σ_ϵ^2 followed by the variances of the random effects in the model. For M_{SDID}, $\boldsymbol{\sigma}$ is defined as $(\sigma_\epsilon^2, \sigma_s^2, \sigma_{s*d}^2, \sigma_i^2, \sigma_{i*d}^2)$. For each model it is assumed that $\boldsymbol{\sigma} \sim Dirichlet(\alpha\mathbf{1})$. The common hyperparameter α means that *a priori* each variance parameter in $\boldsymbol{\sigma}$ has the same marginal distribution. The Dirichlet prior requires that these variance parameters must sum to 1 so that the scale of the latent variables will be established parsimoniously.

The exact value of α is selected so that there is a small chance any given element of $\boldsymbol{\sigma}$ is well below the average value of the variance parameters, the idea being that if one of the variance parameters is atypically small then the corresponding random effect is not important for the model and could be excluded. The specific rule we used for the primate cognitive application was to choose α so that $Pr(\sigma_\theta^2 < E(\sigma_\theta^2)/25) = 0.01$. Note that this rule implies that the value of α depends only on the number of parameters in $\boldsymbol{\sigma}$ because the expected value of each element is the inverse of the number of elements. For the model with 5 variance parameters in $\boldsymbol{\sigma}$, $\alpha = 1.425$.

We assume all of the random effects are conditionally independent and have the following prior distributions: $\boldsymbol{u_s}|\boldsymbol{\sigma} \sim N(\mathbf{0}, \sigma_s^2\mathbf{I})$; $\boldsymbol{u_i}|\boldsymbol{\sigma} \sim N(\mathbf{0}, \sigma_i^2\mathbf{I})$; $\boldsymbol{u_{s,d}}|\boldsymbol{\sigma} \sim N(\mathbf{0}, \sigma_{s*d}^2\mathbf{I})$; $\boldsymbol{u_{i,d}}|\boldsymbol{\sigma} \sim N(\mathbf{0}, \sigma_{i*d}^2\mathbf{I})$. Based on the model fit, we chose an alternative model by removing effects with small variance components. The prior distribution for effects that are excluded from the model are assumed to be degenerate at $\mathbf{0}$.

The final step in setting up the models was to choose the priors for $\boldsymbol{\kappa}$ and $\boldsymbol{\tau}$. The priors selected were the same across models. The κ_j's for rank assessments were assumed to be mutually independent and to have a $Gamma(1,1)$ distribution, which is vague enough that it is compatible with assessments having very many or very few ties. The τ_j's for binomial assessments were assumed to be mutually independent and to have a $Cauchy(0, 0.5^2)$ distribution. This Cauchy distribution is sufficiently vague to comply with assessments being very easy or very difficult, but it also gives considerable prior support to values throughout (0,1). Another viable option would be to assume each τ_j has a standard normal distribution. Such a choice induces a uniform (0,1) prior on the marginal probability of success because of how we selected the prior distribution for the latent variables $\mathbf{z}$. While our application of the model uses the Cauchy prior distribution for the τ_j's, we also considered the standard normal option in our sensitivity and robustness analyses.

The MCMC algorithm was run using 450K burn-in iterations (during the first 200K of which the tuning parameters for the MCMC algorithm were allowed to be adjusted). After this burn-in period, 5 million more iterations were performed. For a more efficient MCMC implementation, we did not sample the missing y_{ijt} values; only the observed y_{ijt} affect the likelihood. Likewise, we ignored any random effects that were only related to missing responses.

3.2 Model Comparison Metrics

We fitted the joint binomial/rank-response data model to the primate cognition data while allowing for species, individual, species*domain, and individual*domain random effects. After doing so, the random effects with small variance components were omitted, forming a simpler alternative. In order to assess the relative performance of each model, the iterations from the MCMC algorithm were used to first obtain draws of the latent variables z_{ijt} and then obtain simulated response variables. Specifically, thinned posterior draws for the random effects and for the error variance were used to simulate $\mathbf{z}$. Then using the corresponding posterior draws for τ and $\boldsymbol{\kappa}$, a set of simulated $\mathbf{y}$ values was obtained. This was done K times, with $K = 5000$ for each model.

The posterior predictive draws under each model were compared to the observed data. For each assessment with a binomial response, the average absolute deviation between the actual and simulated counts was computed across animals. From the kth posterior predictive simulation, the average absolute deviation for assessment j, or $AAD_k(j)$, was computed as

$$AAD_k(j) = \frac{1}{n_j} \sum_i |y_{ij} - y_{ij}^{(k)}|$$

where n_j represents the number of animals with an observed response for assessment j, y_{ij} represents the total number of successes for individual i in assessment j, and $y_{ij}^{(k)}$ represents the posterior-predictive simulated value for y_{ij} in the k^{th} simulation, $k = 1, \ldots, K$.

For each assessment with a rank response, the proportion of pairs for which the simulated y's and actual y's were completely concordant was computed; completely concordant implies that the higher actual rank in a given pair also had the higher simulated rank, or if the actual ranks were tied the simulated ranks were also tied. From the kth posterior predictive simulation, this proportion of completely concordant pairs for assessment j, or $PropCC_k(j)$, was computed as

$$\begin{aligned} PropCC_k(j) &= \frac{1}{n_j(n_j-1)/2} \sum_i \sum_{i'<i} \\ &\quad \times 1(sign(y_{ij} - y_{i'j}) = sign(y_{ij}^{(k)} - y_{i'j}^{(k)})). \end{aligned}$$

From the 5000 posterior predictive simulated data sets, the median of the $AAD_k(j)$ and $PropCC_k(j)$ values were computed to provide a central measure of the criteria; the 5th and 95th percentiles were also computed to reflect the spread of the criteria. Smaller values are desirable for $AAD(j)$

while larger values are desirable for $PropCC(j)$.

4 Results

Section 4.1 compares results from two models considered—one with four random effects and a reduced model with only two random effects. The models are compared in terms of their variance parameter estimates and how well their posterior predictive distributions match the observed data. One model is chosen as the preferred model. Section 4.2 contains more detailed results from the preferred model.

4.1 Model Comparison

The initial model for primate cognition, containing random effects for species, species*domain, individual, and individual*domain, was fitted and examined to identify alternative models. Posterior estimates of each variance parameter are contained in Table 3. The large estimate of σ^2_ϵ reveals that most variability in the latent variables is unexplained by any of the aforementioned effects. Neither individual nor individual*domain effects appear to be very important, while species and species*domain effects do appear to be important. We therefore also posited an alternative model that included only the species and species*domain effects; this model is denoted by M_{SD} in Table 3. This model produces estimates of the variance components that are similar to estimates obtained from the full model, for components that

Table 3: Posterior estimates of variance parameters. Modified from Amici et al. (2012)

Model	Posterior Mean (sd)				
	Species, σ^2_s	Species*Domain, σ^2_{s*d}	Individual, σ^2_i	Individual*Domain, σ^2_{i*d}	Error, σ^2_ϵ
M_{SDID}	0.15 (0.08)	0.16 (0.05)	0.01 ($<$ 0.01)	0.02 (0.01)	0.67 (0.07)
M_{SD}	0.17 (0.09)	0.17 (0.05)			0.66 (0.08)

are common to both models.

Table 4 gives percentiles of the average absolute deviation for each assessment with a binomial assessment. Table 5 gives percentiles of the proportion of completely concordant pairs for each assessment with a rank response.

Not surprisingly, $AAD(j)$ tended to increase with the number of trials in the assessment.The notable exception was assessment 5, for which success was very rare in the observed and simulated data.

Both M_{SDID} and M_{SD} perform similarly in terms of prediction. The 5th and 95th percentiles of the average absolute deviations and the concordant-pair proportions convey some sense of the criteria's variability and generally exhibit substantial overlap between the two models. The agreement between the data and the posterior predictive distributions is only poor for the second (rank) assessment. This assessment is characterized by a very large proportion of ties in the rank data. As a secondary measure, the proportion of completely discordant pairs in assessment 2 (i.e., $y_{i2} > y_{i'2}$ in the observed data but $y_{i2} < y_{i'2}$ in the posterior predictive data, or vice-versa) ranged from 0.08 to 0.12 across the models.

Because both models provided approximately the same fit to the data, we

Table 4: Median (and 5th, 95th percentiles) of average absolute deviation between observed and 5000 posterior-predictive simulated responses for the binomial assessments

Assessment (# of Trials)	Median (and 5th, 95th Percentiles) of Average Absolute Deviation M_{SDID}	M_{SD}
3 (2)	0.48 (0.36, 0.66)	0.60 (0.47, 0.74)
5 (10)	0.61 (0.46, 0.76)	0.75 (0.58, 0.93)
6 (3)	0.64 (0.51, 0.78)	0.67 (0.54, 0.79)
7 (3)	0.90 (0.73, 1.10)	0.84 (0.71, 0.98)
8 (2)	0.51 (0.41, 0.62)	0.56 (0.44, 0.67)
9 (2)	0.61 (0.49, 0.70)	0.64 (0.51, 0.75)
10 (1)	0.34 (0.28, 0.44)	0.39 (0.30, 0.49)
11 (1)	0.11 (0.11, 0.23)	0.20 (0.13, 0.30)
12 (6)	1.35 (1.04, 1.84)	1.16 (0.99, 1.35)
13 (6)	1.49 (1.14, 2.36)	1.25 (1.07, 1.43))
14 (6)	1.42 (1.13, 1.91)	1.26 (1.09, 1.45)
15 (6)	1.41 (1.13, 1.94)	1.31 (1.13, 1.50)
16 (6)	1.54 (1.21, 2.28)	1.32 (1.13, 1.51)
17 (6)	1.43 (1.15, 2.06)	1.29 (1.12, 1.49))

Table 5: Proportion of pairwise individual comparisons that were completely concordant between observed and posterior predictive simulated responses for the rank assessments. The median (5th, 95th percentiles) of the 5000 iteration-specific concordance proportions provide a point estimate (information about the spread)

Assessment	Median (and 5th, 95th Percentiles) of Completely Concordant Pairs Proportion M_{SDID}	M_{SD}
1	0.61 (0.35, 0.75)	0.51 (0.30, 0.75)
2	0.36 (0.33, 0.45)	0.35 (0.32, 0.42)
4	0.68 (0.58, 0.72)	0.60 (0.54, 0.66)

chose as our preferred model M_{SD}, the model with species and species*domain effects but without individual and individual*domain effects. However, it is important to note that both the reduced model and full model lead to the same qualitative conclusions regarding the substantive questions posed in this analysis. In particular, domain*species effects are important in predicting outcomes in both models, and individual*domain effects were necessary in the model that included individual effects. Thus, assessment domain seems to be an important factor in explaining the relative performance of subjects on the test.

4.2 Results from Preferred Model

The selected model, M_{SD}, has random effects for both species and species*domain effects, so it is natural to examine the relative importance of these two effects. Figure 1 displays the posterior distribution of these variance parameters.Note that σ_ϵ^2 is not depicted since its value is implicit from the prior constraint.

Posterior inference on $\boldsymbol{\sigma}$ suggests that the species and species*domain effects might have similar variability. The posterior mean (standard deviation) of σ_θ^2 is 0.170 (0.089) and of σ_ω^2 is 0.166 (0.055). However, the error variability tends to be the primary source of variability in $\mathbf{z}$, accounting for roughly two-thirds of the total variance in the latent variables.

Besides the variance parameters, other model quantities of note are the random effects. Figure 2 depicts the estimated marginal posterior distributions of the species main effects and of the species*domain interaction effects.

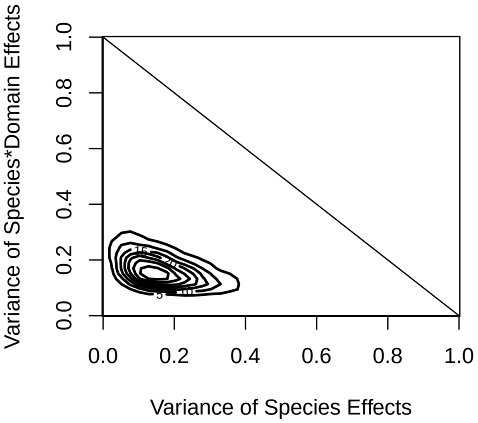

Figure 1: Posterior distribution of the variance parameters σ_θ^2 (for species effects), σ_ω^2 (for species*domain effects), and $\sigma_\epsilon^2 \equiv 1 - \sigma_\theta^2 - \sigma_\omega^2$ (for error terms). The posterior density is zero outside of the demarcated triangular region.

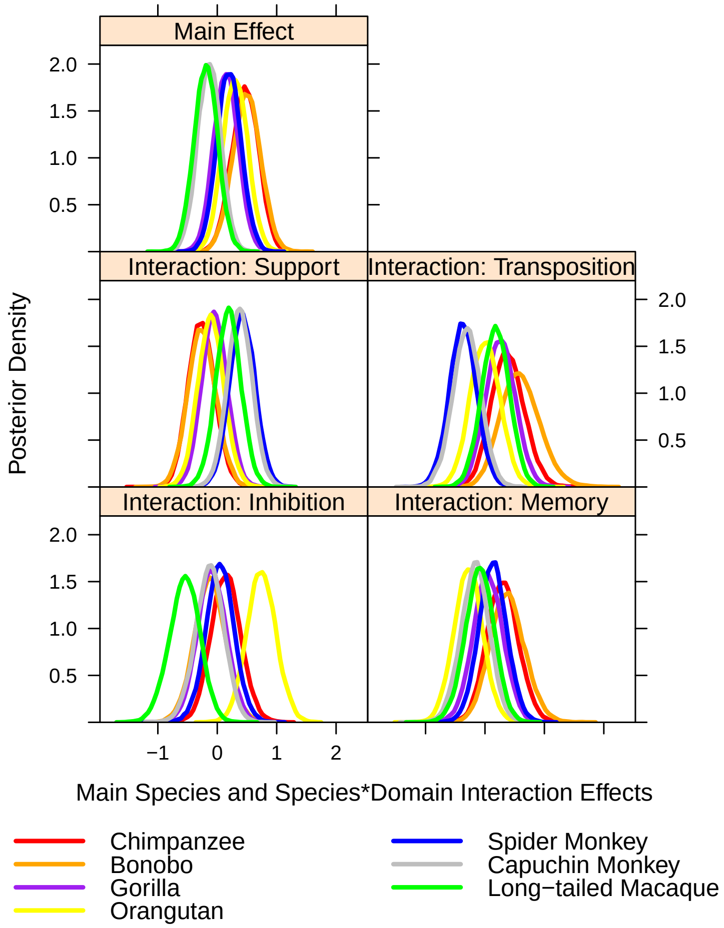

Figure 2: Estimated marginal posterior distributions of the species main effects (top plot) and of the species*domain interaction effects (bottom four plots). Modified from Amici et al. (2012).

The importance of the species*domain interaction effect is illustrated in Figure 2. In this plot, orangutan performance in the inhibition domain is much better relative to the other species than might have been expected based on the other domains. Similarly, spider monkeys excelled relative to the other species in the support domain, but were generally average in other domains.

Figure 3 depicts the posterior probability that one species performs better in a given domain than another species does. This figure is necessary to properly compare species performance because of the correlations between species and species*domain effects in the posterior distribution.

Several interesting findings are apparent in Figure 3. As a group the great apes performed better than the monkeys in all domains except the support domain. The spider monkeys tended to outperform the other monkey species. Other than exceptionally good (poor) performances by spider monkeys (long-tailed macaques), there were not great differences in performance in the support domain. Another notable finding is the similarity of chimpanzees and bonobos in each of the four domains.

4.3 Robustness and Sensitivity Analyses

To increase confidence in our findings, we investigated the model fit and conducted a number of sensitivity and robustness analyses. To assess model fit, we performed a number of diagnostics based on examining the distribution of the latent traits. In general, these diagnostics suggested no model lack-of-

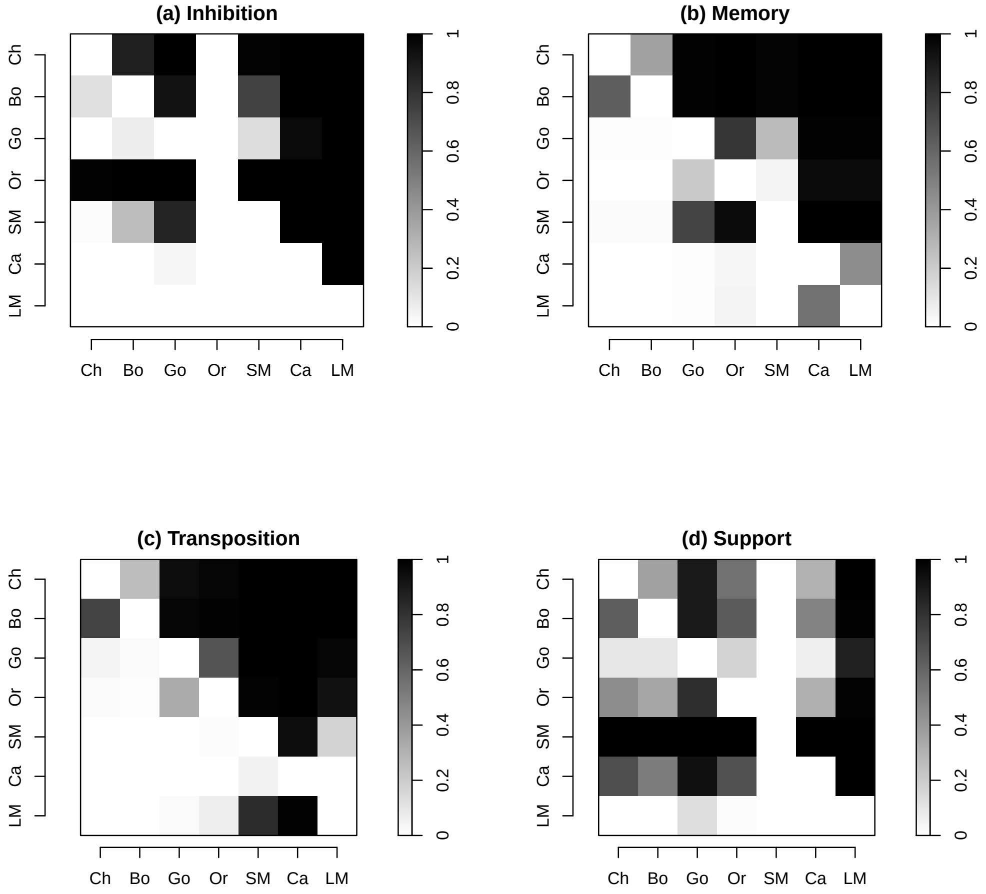

Figure 3: Estimated posterior probability that the row species had higher cognitive ability than the column species, within each of the four domains. Modified from Amici et al. (2012).

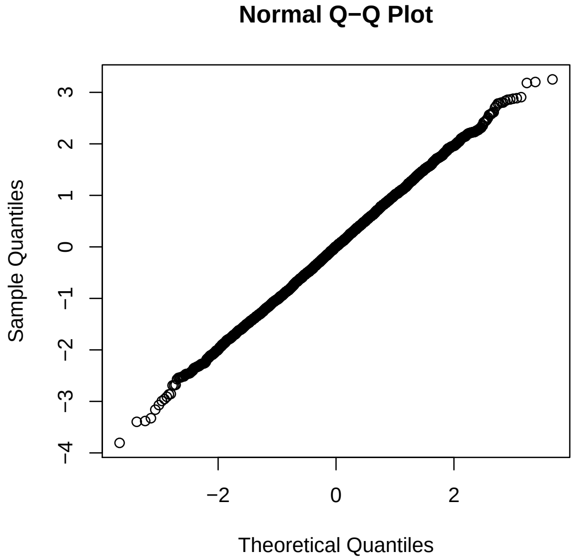

Figure 4: Normal probability plot of standardized residuals from M_{SD} (last iteration of MCMC algorithm).

fit. For example, a normal probability plot of the standardized residuals from the last MCMC iteration of this model is depicted in Figure 4. Because M_{SD} assumes that each $z_{ijt}|u_{s(i)}, u_{s(i),d(j)}, \sigma^2_\epsilon \overset{ind}{\sim} N(u_{s(i)} + u_{s(i),d(j)}, \sigma^2_\epsilon)$, it follows that each standardized residual

$$r_{ijt} = \frac{z_{ijt} - u_{s(i)} - u_{s(i),d(j)}}{\sigma_\epsilon}$$

should be independently distributed according to a standard normal distribution (Johnson, 2007). Figure 4 appears consistent with this assumption.

Along with assessing the model fit for our selected model, we conducted a number of sensitivity analyses by altering the prior distribution. One alteration was to replace the $Cauchy(0, scale = 0.5)$ distribution with the standard normal distribution for the prior of each threshold parameter (τ_j), which implies that the marginal prior probability of success is uniformly distributed. It is also natural to wonder how the variance components are affected by changes to the prior distribution. We assumed the prior distribution of $\boldsymbol{\sigma}$ was $Dirichlet(\alpha\mathbf{1})$, where α was selected to ensure a small prior probability that any element of $\boldsymbol{\sigma}$ was much smaller than the others. This had been motivated by a desire to have a non-local prior density, as defined in (Johnson and Rossell, 2010). We selected a smaller value of α to ascertain its effect on the variance components.

Because the data were collected by various research teams, it is natural to wonder if there is a researcher effect on performance. This proves difficult to assess because there is an inherent confounding between species*domain effects and researcher effects: only one research team collected data from the three monkey species, and this same research team was the only one that collected data from assessments in multiple domains. By excluding the data from great apes, all remaining data are from the same research team (Amici's), and so we applied the model to only the data from monkeys to determine if the results were materially different in a situation where the research team was unchanged.

Finally, we investigated how the joint analysis of binomial and rank responses differed from the analysis of only binomial outcomes. We note that because there were so few rank-response outcomes, it was not feasible to estimate the rank–species*domain interactions based only on the rank data. By comparing inferences from the combined data and the binomial data, we were able to implicitly assess the impact of the rank data on our conclusions.

The results of our sensitivity analyses are summarized in Table 6. There is much support for the qualitative finding that the individual and individual*domain effects explain very little of the latent performance. The species effects have relatively less importance when the prior distribution on the variance parameters is vague; as expected, its variance component is most affected because there are only seven distinct species, and thus relatively little information from the data to update this parameter. The importance of the species*domain interaction effects is very consistent, except it is notably higher when the rank data are excluded. The ability to jointly model mixed data types is not only preferable because it makes use of all available data, but it also leads to differences in the estimated magnitude of various model effects. If the threshold parameters $\boldsymbol{\tau}$ are modeled with a standard normal distribution, there is only a slight impact on the variance components: the species effect becomes less pronounced.

5 Discussion

We have presented a framework for analysis of data from multiple assessments, some with binomial responses and others with rank responses. To do so, we used data augmentation, employing a random effects model for the

Table 6: Posterior estimates of variance parameters under various changes to data used, prior distribution

Model	Change	Posterior Mean				
		Sp., σ_s^2	Sp.*Dom., σ_{s*d}^2	Ind., σ_i^2	Ind.*Dom., σ_{i*d}^2	Error, σ_ϵ^2
M_{SDID}	Original	0.15	0.16	0.01	0.02	0.67
M_{SDID}	Vague Prior	0.06	0.19	0.00	0.01	0.73
M_{SDID}	$\tau_j \sim N(0,1)$	0.12	0.16	0.01	0.02	0.69
M_{SDID}	Monkey Only	0.12	0.12	0.01	0.01	0.75
M_{SDID}	Binomial Only	0.14	0.24	0.01	0.01	0.60
M_{SD}	Original	0.17	0.17			0.66
M_{SD}	Vague Prior	0.09	0.18			0.74
M_{SD}	$\tau_j \sim N(0,1)$	0.14	0.17			0.69
M_{SD}	Monkey Only	0.14	0.13			0.73
M_{SD}	Binomial Only	0.16	0.24			0.60

latent variables. We also used Dirichlet priors to establish the scale for the latent variables. We outlined a Metropolis-within-Gibbs MCMC algorithm that can be used for posterior sampling, with an innovation that enables the Markov chain to mix better by allowing for simultaneous shifts in latent variables associated with tied rankings.

The application of this model to primate cognition data allowed us to address a number of important scientific issues. In particular, we were able to study the relative importance of various cognitive effects in analyzing cognitive performance of non-human primates in assessments from four domains. The key findings are that there are substantial species main effects and important species*domain interaction effects. That is, the relative cognitive performance of species varies systematically across primate species, but this variability cannot be explained solely by a general intelligence latent trait. Our conclusion that different domains tend to yield different ordering of species performance is consistent with findings of other studies (see Tomasello and Call 1997, and Shettleworth 2010 for a review.)

An in-depth discussion of our findings in light of other research on non-human primates was previously provided by Amici et al. (2012). However, we note that Johnson et al. (2002) did not find strong evidence of genus*domain interactions in their meta-analysis of primate cognition (though genus effects were included), although Banerjee et al. (2009), using rank data analyses similar to those proposed in Johnson et al. (2002), did report evidence of genus*domain interactions. While it is difficult to identify the exact nature of these seemingly contradictory finding, there are several possible explanations. Among these are differences in assessments, domains, and species across the studies. Another possible explanation is that the current data set is much more extensive and contains many assessments using binomial responses. Furthermore, many of the binomial assessments were based on six or more trials, which provide higher quality data for detecting species*domain interactions. Most importantly, however, is the manner in which data were collected. Data reported in Johnson et al. (2002) were based on an historical meta-analysis of data collected over 70 years from multiple researchers, whereas data analyzed here and in Banerjee et al. (2009) were collected under more controlled conditions.

In this study, we found substantially more systematic interspecific variation than systematic intraspecific variation. This finding may not hold in studies with greater variation in individual age, different assessments and domains, or different species. Care must be taken in interpreting the results. The unbalanced nature of the data (not every individual was measured for each assessment) raises concerns over a possible confounding of interaction terms with the experimental design. Similarly, the number of animals ranked in a particular assessment might have an impact on the model conclusions in two regards: (1) as more animals are ranked, the ordering becomes more informative; and (2) there are subtle changes to the implied probability of tied rankings at the extremes relative to the interior, based on the behavior of the order statistics as sample size increases. Finally, there is the possibility of experimenter effects because different researchers collected the data; however, the conclusions were similar when only data from the monkey species were used (all of which was collected by the same researcher).

The techniques we described and their application to primate cognitive testing represent an important step forward. Not only does the modeling approach allow inference using both binomial and rank data, but it also emphasizes parsimony by constraining the prior distribution of the latent variables in $\mathbf{z}$ to have a common marginal mean (0) and marginal variance (1), regardless of the number of random factor effects that are included. This consistency is desirable to simplify prior elicitation, interpretation of random effects, and comparisons across models.

There are other applications in which the joint model may be particularly helpful. For example, a marketing firm may administer a survey wherein consumers rank items based on preference or propensity to purchase, providing responses that are intrinsically on a rank scale. They could supplement these preferences with actual purchasing behavior, such as whether or not the item had been purchased in the previous six weeks, and thus obtain binomial data as well. We are currently examining the extension of this model framework to consumer choice modeling and purchasing patterns, and feel that the model proposed here for comparing cognitive abilities of non-human primates will find much broader application in marketing research.

The scope of potential model application is greatly enlarged by recognizing that not only are data frequently recorded as ranked values, but data are often converted to ranks to provide more robust analyses. This conversion from the original measurements to rank data is especially helpful in situations where the likelihood of the data is unknown or when outliers are particularly influential.

While not the focus of our work, the model may be expanded to include even more response types, such as ordinal, count, or continuous data. Inasmuch as the various response types may have their likelihood expressed in terms of underlying latent variables with a standard normal distribution, the proposed model framework may be used. For example, Fahrmeir and Raach (2007) note that if one wants to model normally distributed data, then the latent variables may be assumed to coincide with the observed data; for the identification restrictions we espouse, the observed data would instead be viewed as a location-scale transformation of the latent standard normal variables. Another increasingly common technique for jointly modeling different response types is to model dependence with a Gaussian copula and let the marginal distributions of each response type be modeled appropriately. Future research will investigate how such models may be adapted to explicitly model the probability of ties in ranked data.

References

Albert, J. H. and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679.

Amici, F., Barney, B., Johnson, V. E., Call, J., and Aureli, F. (2012), "A Modular Mind? A Test Using Individual Data from Seven Primate Species," *PLoS ONE*, 7, e51918.

Banerjee, K., Chabris, C., Johnson, V., Lee, J., Tsao, F., and Hauser, M. (2009), "General Intelligence in Another Primate: Individual Differences across Cognitive Task Performance in a New World Monkey (*Saguinus Oedipus*)," *PLoS ONE*, 4, e5883.

Cowles, M. K. (1996), "Accelerating Monte Carlo Markov Chain Convergence for Cumulative-link Generalized Linear Models," *Statistics and Computing*, 6, 101–111.

Dunson, D. B. (2000), "Bayesian Latent Variable Models for Clustered Mixed

Outcomes," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 62, 355–366.

Fahrmeir, L. and Raach, A. (2007), "A Bayesian Semiparametric Latent Variable Model for Mixed Responses," *Psychometrika*, 72, 327–346.

Gruhl, J., Erosheva, E. A., and Crane, P. K. (2013), "A Semiparametric Approach to Mixed Outcome Latent Variable Models: Estimating the Association between Cognition and Regional Brain Volumes," *The Annals of Applied Statistics*, 7, 2361–2383.

Gueorguieva, R. V. and Agresti, A. (2001), "A Correlated Probit Model for Joint Modeling of Clustered Binary and Continuous Responses," *Journal of the American Statistical Association*, 96, 1102–1112.

Hoff, P. D. (2007), "Extending the Rank Likelihood for Semiparametric Copula Estimation," *The Annals of Applied Statistics*, 1, 265–283.

Johnson, V. E. (2007), "Bayesian Model Assessment Using Pivotal Quantities," *Bayesian Analysis*, 2, 719–734.

Johnson, V. E., Deaner, R. O., and van Schaik, C. P. (2002), "Bayesian Analysis of Rank Data With Application to Primate Intelligence Experiments," *Journal of the American Statistical Association*, 97, 8–17.

Johnson, V. E. and Rossell, D. (2010), "On the Use of Non-local Prior Densities in Bayesian Hypothesis Tests," *Journal of the Royal Statistical Society,* Ser. B, 72, 143–170.

Murray, J. S., Dunson, D. B., Carin, L., and Lucas, J. E. (2013), "Bayesian Gaussian Copula Factor Models for Mixed Data," *Journal of the American Statistical Association*, 108, 656–665.

Quinn, K. M. (2004), "Bayesian Factor Analysis for Mixed Ordinal and Continuous Responses," *Political Analysis*, 12, 338–353.

Shettleworth, S. J. (2010), *Cognition, evolution, and behavior*, Oxford: Oxford University Press.

Tanner, M. A. and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82, 528–540.

Tomasello, M. and Call, J. (1997), *Primate Cognition*, Oxford: Oxford University Press.