BAYESIAN JOINT MODELING OF BINOMIAL AND RANK RESPONSE DATA

A Dissertation

by

BRADLEY JOHN BARNEY

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2011

Major Subject: Statistics

BAYESIAN JOINT MODELING OF BINOMIAL AND RANK RESPONSE DATA

A Dissertation

by

BRADLEY JOHN BARNEY

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Co-Chairs of Committee,	Valen E. Johnson
	Simon J. Sheather
Committee Members,	Veerabhadran Baladandayuthapani
	Dudley L. Poston, Jr.
Head of Department,	Simon J. Sheather

August 2011

Major Subject: Statistics

ABSTRACT

Bayesian Joint Modeling of Binomial and Rank Response Data. (August 2011)

Bradley John Barney, B.A., Brigham Young University;

M.S., Brigham Young University

Co–Chairs of Advisory Committee: Dr. Valen E. Johnson
Dr. Simon J. Sheather

We present techniques for joint modeling of binomial and rank response data using the Bayesian paradigm for inference. The motivating application consists of results from a series of assessments on several primate species. Among 20 assessments representing 6 paradigms, 6 assessments are considered to produce a rank response and the remaining 14 are considered to have a binomial response. In order to model each of the 20 assessments simultaneously, we use the popular technique of data augmentation so that the observed responses are based on latent variables. The modeling uses Bayesian techniques for modeling the latent variables using random effects models. Competing models are specified in a consistent fashion which easily allows comparisons across assessments and across models. Non-local priors are readily admitted to enable more effective testing of random effects should Bayes factors be used for model comparison. The model is also extended to allow assessment-specific conditional error variances for the latent variables. Due to potential difficulties in calculating Bayes factors, discrepancy measures based on pivotal quantities are adapted to test for the presence of random effects and for the need to allow assessment-specific conditional error variances. In order to facilitate implementation, we describe in detail the joint prior distribution and a Markov chain Monte Carlo (MCMC) algorithm for posterior sampling. Results from the primate intelligence data are presented to illus-

trate the methodology. The results indicate substantial paradigm-specific differences between species. These differences are supported by the discrepancy measures as well as model posterior summaries. Furthermore, the results suggest that meaningful and parsimonious inferences can be made using the proposed techniques and that the discrepancy measures can effectively differentiate between necessary and unnecessary random effects. The contributions should be particularly useful when binomial and rank data are to be jointly analyzed in a parsimonious fashion.

To Tammy, Tyson, Kevin, and Matthew

ACKNOWLEDGMENTS

There are so many people who have had a hand in my education, and I would like to briefly give special mention to those individuals who have been especially influential.

My parents instilled in me a love of learning and gave me support in countless ways to pursue that knowledge. My wife's parents have also been very supportive of my decision to pursue a Ph.D. My wife has encouraged me and supported me, especially when times were most challenging. Her many sacrifices enabled me to focus on my studies. I would also like to thank each of my children, because when I came home from a day filled with setbacks and dead ends they helped me to remember that there is more to life than statistics.

I would like to thank Federica Amici for her collaboration on the primate intelligence application and for sharing her data. In addition, Federica and I wish to thank the following individuals for generously allowing use of their primate intelligence data: Filippo Aureli, Josep Call, Brian Hare, Marc Hauser, Esther Herrmann, Alexandra Rosati, Jeffrey Stevens, Elisabetta Visalberghi, Petra Vlamings, and Victoria Wobber. I would like to thank Texas A&M University and the Department of Symptom Research at the University of Texas M. D. Anderson Cancer Center for directly and indirectly providing financial assistance. I also appreciate the assistance of Clift Norris in my attempts to increase the efficiency of the computer programs used to fit the various primate intelligence models considered.

I would like to thank the many professors at Eastern Arizona College, Brigham Young University, and Texas A&M University who patiently answered my repeated questions and encouraged my progress. Three individuals in particular have changed the course of my life for the better. Dr. Michael Speed was crucial in my transition to

Texas A&M and I am especially grateful for how he looked out for my best interests and gave me multiple opportunities to succeed, all the while teaching me through his example how to live and how to work. Dr. Simon Sheather not only co-chaired my graduate committee but also gave me opportunities to collaborate on very interesting research. His words of praise always lifted me up and made me want to merit his approval. I owe a tremendous amount of gratitude to Dr. Valen Johnson. I simply could not have completed this work without his aid. I deeply admire his brilliance, innovation, and practicality. It has been an honor to work under his tutelage as he has guided me through this work.

In addition to Dr. Sheather and Dr. Johnson, I would like to thank my other committee members, Dr. Veerabhadran Baladandayuthapani and Dr. Dudley L. Poston, Jr., for their service.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

FIGURE Page

FIGURE Page

CHAPTER I

INTRODUCTION

There are many types of response data, and each has associated modeling methodology with varying levels of sophistication. Binomial data are very common, which might explain the abundance of available approaches for modeling such data. Rank response data are also quite common, especially when considering the frequency with which raw response data are transformed to represent a rank; several nonparametric methods employ a rank transformation. Sometimes a multivariate response vector contains different types of responses—some data might be continuous, some count data, some binomial data, etc. In such instances, it is essential that the modeling techniques are flexible enough to model each response appropriately.

We present techniques for joint modeling of binomial and rank response data using the Bayesian paradigm for inference. The motivating application consists of results from a series of assessments on several primate species. Among 20 assessments, 6 are considered to produce a rank response and the remaining 14 are considered to have a binomial response. In order to model each of the 20 assessments simultaneously, we use the popular technique of data augmentation so that the observed responses are based on latent variables. Not only does this accommodate joint modeling, it does so in a coherent fashion which easily allows comparisons across assessments. We allow the latent variables to be influenced by random effects. Discrepancy measures based on pivotal quantities are used to test for random effects.

The novel modeling uses Bayesian techniques because, as many have argued, they

---

This dissertation follows the style of the *Journal of the American Statistical Association.*

are especially appealing for naturally incorporating uncertainty in prior information and in posterior inference. In order to facilitate implementation, we describe in detail the joint prior distribution and a Markov chain Monte Carlo (MCMC) algorithm for posterior sampling.

The remainder of this dissertation focuses on the following contributions:

- Introduction of a random effects model for joint analysis of rank and binomial data, including a unique approach to ensure model identifiability in a parsimonious fashion that conveniently admits a non-local prior density on selected parameters;
- Details of an MCMC algorithm for posterior sampling;
- Use of an existing discrepancy measure to test for random effects and its adaptation to test for nonconstant variance; and
- Results of the model application to primate intelligence data.

Chapter II consists of a literature review to highlight particularly relevant research. Chapter III introduces the joint model, gives details of the MCMC algorithm, and discusses use of discrepancy measures for hypothesis testing and model comparison. Chapter IV presents results from the primate intelligence application, including model comparison results and sensitivity to modest changes in the prior distribution or the exclusion of some data. Chapter V discusses the contributions made in this paper as well as potential extensions of the methodology and computer program to fit these models.

CHAPTER II

RELEVANT LITERATURE

In this chapter relevant literature and techniques are summarized. The summary is divided into several sections. Section A briefly summarizes a Bayesian technique for analysis of binomial data. Section B highlights a Bayesian technique for analysis of rank data. Section C discusses Bayes factors and discrepancy measures for model selection. Section D explains non-local priors. Finally, Section E discusses literature on intelligence testing and assessment in primates because of the relevance to the desired application.

A. Bayesian Analysis of Binomial Data

There is an abundance of binomial data, and this has likely contributed to a wide variety of available tools for their analysis. One particularly popular approach to analyzing binomial data is logistic regression. Another related approach is probit regression, which uses the standard normal cumulative distribution as the link function in a generalized linear model. In the context of Bayesian analyses, Albert and Chib (1993) made a landmark contribution to probit regression with their use of data augmentation. The idea behind data augmentation, popularized in Bayesian analyses by Tanner and Wong (1987), is to supplement the observed data with latent data and thus form the complete data; its power comes from situations where analysis of the complete data is straightforward but the observed data alone is not. The application of this to probit regression by Albert and Chib (1993) is now described. Let y_i denote a Bernoulli response. It is assumed that y_i has an associated latent variable z_i, which some authors (e.g., Fahrmeir and Raach 2007) also refer to as an underlying variable. It is also assumed that there is a threshold, or cutpoint, parameter τ; Albert and Chib

(1993) assumed $\tau \equiv 0$ but this restriction can be relaxed if an informative prior is placed on each z_i. The relationship between y_i and z_i can be understood by treating y_i as an indicator of whether z_i exceeds the threshold. That is, $y_i = 1(z_i > \tau)$.

The power of this approach can be understood by considering models for z_i. A particularly convenient model is to assume that z_i has a conditionally normal distibution which can depend on covariates. Let $\mathbf{x}_i^T$ be a vector of covariates associated with observation i. Assume $z_i|\mathbf{x}_i^T, \boldsymbol{\beta}, \sigma^2 \sim N(\mathbf{x}_i^T\boldsymbol{\beta}, \sigma^2)$ and that the z_i's are conditionally independent. Then with a normal-inverse gamma prior on $(\boldsymbol{\beta}, \sigma^2)$, all complete conditionals are conveniently sampled using either a normal, truncated normal, or inverse gamma distribution. This makes the posterior distribution especially easy to sample through Gibbs sampling while allowing covariates to affect the probability of a success.

The data augmentation technique for Bernoulli responses can also be adapted for modeling ordinal, nominal, or binomial outcomes. Albert and Chib (1993) presented the extension to ordinal and nominal outcomes with more than two possible responses. Binomial data can be modeled by using multiple Bernoulli variables. For example, y_i successes out of T trials for subject i can be modeled by $y_{it}, t = 1, \ldots, T$. If the result of each trial is available, each y_{it} is known. If only the summary statistic y_i is known, it can be assumed that the first y_i Bernoulli variables equal 1 and the remaining $T - y_i$ Bernoulli variables equal 0. This is justifiable because binomial data arise as the sum of independent and identically distributed Bernoulli random variables, so without loss of generality it can be assumed successes occurred first.

Joint modeling of ordinal and continuous data has been done with data augmentation (e.g., Quinn 2004; Fahrmeir and Raach 2007). The latent variables for continuous data are redundant in that they equal the observed value, and they do not have associated thresholds but do have free variances. The latent variables for

ordinal data with K categories have constrained variances and $K-1$ free thresholds.

B. Bayesian Analysis of Rank Data

Rank data typically arise in one of two situations. The first situation occurs when only an ordering (or partial ordering) can be made without any quantitative measurement as to the degree of superiority. The second situation occurs when there are available quantitative measurements, but because of other considerations it is preferred to use the rank transformation. For example, nonparametric procedures often employ a rank transformation to increase model robustness by reducing the potential impact of outliers.

Some classical techniques for analysis of rank data were reviewed by Marden (1995). Marden devoted an entire chapter to a summary of considerations for data with ties, partial orderings, and incomplete rankings (ch. 11). Such possibilities can greatly complicate what might otherwise be a simple rank-based analysis (Johnson et al. 2002, p. 16).

Johnson et al. (2002) applied the data augmentation strategy to develop a model for analyzing rank data while allowing for explicit modeling of the probability that two given observations are tied. This approach, using the Bayesian perspective, is now restated with some changes to notation in the simplified case where data are only available from one assessment.

Suppose that I subjects are ranked based on an assessment, with y_i denoting the rank of the i^{th} subject, $i = 1, \ldots, I$. Unlike Johnson et al. (2002), assume that a higher value of the rank denotes better (rather than worse) performance. Regardless of this change, assume that each observed rank variable y_i has an associated latent variable z_i. If no ties are permitted in rankings, $\mathbf{y}$ and $\mathbf{z}$ are linked through the

condition that $y_i < y_{i'} \Leftrightarrow z_i < z_{i'}\ \forall i \neq i'$. Thus, the ordering of the latent variables corresponds to the ordering of the observed ranks. As with Bernoulli data, the latent variables $\mathbf{z}$ can be modeled as conditionally independent and conditionally normally distributed.

When ties for observed ranks are permitted, Johnson et al. (2002) explicitly modeled the probability that two observations are tied based on the scaled distance between the latent variables, with the scale determined by a parameter κ. Specifically, they assumed that

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

and that

$$Pr(y_k < y_{k'}|z_k, z_{k'}, \kappa) = (1 - Pr(y_k = y_{k'}|z_k, z_{k'}, \kappa))1(z_k < z_{k'}).$$

These formulas are used to define the likelihood of an observed set of rankings based on quantities of the form $p_{(i)}(\kappa)$, $i = 1, \ldots, I-1$, defined as

$$p_{(i)}(\kappa) = \begin{cases} \exp(-(z_{(i+1)} - z_{(i)})/\kappa), & \text{if } y_{(i)} = y_{(i+1)} \\ 1 - \exp(-(z_{(i+1)} - z_{(i)})/\kappa), & \text{if } y_{(i)} < y_{(i+1)} \end{cases} \tag{2.1}$$

where $y_{(i)}$ ($z_{(i)}$) is the i^{th} smallest observed value (latent variable) for $i = 1, \ldots, I$ (see Johnson et al. 2002, eq. 2).

Then the entire likelihood can be summarized as

$$f(\mathbf{y}|\mathbf{z}, \kappa) = \prod_{i=1}^{I-1} \left[p_{(i)}(\kappa) \prod_{i'=i+1}^{I} 1(\{z_{i'} \leq z_i \cap y_{i'} \leq y_i\} \cup \{z_{i'} \geq z_i \cap y_{i'} \geq y_i\}) \right] \tag{2.2}$$

(compare with Johnson et al. 2002, eq. 3). The inner product is necessary to ensure that the ordering of the observed variables is consistent with the ordering of the latent

variables.

This work can be extended by jointly modeling binomial and rank data, a pursuit of particular importance in the motivating application: primate intelligence. In addition, changes to conventional techniques for modeling of the latent variables $\mathbf{z}$, specifically in how their scale is established, can be made to allow harmonious model comparison. Generally in a Bayesian probit regression analysis it is assumed that the variance of the error terms, conditional on the fixed effects, is fixed at 1. Johnson et al. (2002) established scale in their analysis of rank data by fixing the variance for a particular effect at 1. Either of these approaches leads to different marginal variability in models with different numbers of random effects. The proposed joint analysis establishes scale in such a manner that the marginal variability of $\mathbf{z}$ is 1, regardless of the number of random effects. See Chapter III for further details.

C. Bayes Factors and Discrepancy Measures

Kass and Raftery (1995) described the use of Bayes factors in performing Bayesian hypothesis tests. The Bayes factor for comparing two models is equal to the ratio of marginal likelihoods under each model. As is widely known, the marginal likelihood can be difficult if not impossible to calculate because it requires integration of the likelihood with respect to the prior distribution on model parameters. The integral might not have a closed form or might be highly dimensional, and therefore many methods have been proposed to approximate Bayes factors or variations of Bayes factors such as the pseudo-Bayes factor (see, e.g., Gelfand and Dey 1994; Kass and Raftery 1995). Lopes and West (2004) compared many techniques in a simulation study involving selection of the number of factors in a factor analysis. One of the techniques that performed well in their study can be useful in many classes of mod-

els. This approximation to the marginal likelihood, due to Gelfand and Dey (1994), uses the technique of importance sampling with draws coming from the posterior distribution.

Specifically, let $\boldsymbol{\theta_A}$ represent all of the parameters in model M_A, and let $f(\mathbf{y}|\boldsymbol{\theta_A})$ represent the likelihood under M_A. Gelfand and Dey (1994) noted that for an arbitrary density $h(\boldsymbol{\theta_A})$ with the same support as the posterior distribution, $\pi(\boldsymbol{\theta_A}|\mathbf{y})$, the marginal likelihood can be expressed as the inverse of

$$\int_{\boldsymbol{\Theta_A}} \frac{h(\theta_A)}{f(y|\theta_A)\pi(\theta_A))}\pi(\theta_A|y)\mathrm{d}\boldsymbol{\theta_A}$$

(p. 511). Using B draws from the posterior distribution, say $\theta_{A,1}, \ldots, \theta_{A,B}$ (which might be obtained via MCMC), the marginal likelihood can be estimated by

$$\left(\frac{1}{B}\sum_{b=1}^{B}\frac{h(\theta_{A,b})}{f(y|\theta_{A,b})\pi(\theta_{A,b})}\right)^{-1} \tag{2.3}$$

(eq. 27).

Applying the approximation in Equation 2.3 to each model under consideration yields an estimated marginal likelihood for each model. The Bayes factors comparing any two models are then readily estimated by the ratio of estimated marginal likelihoods.

However, the approach using Equation 2.3 is not easily used when latent variables have been introduced (Chib 2001, p. 3627). Suppose that two models for data $\mathbf{y}$, M_A and M_B, each involve $\mathbf{z}$, where $\mathbf{z}$ consist of latent variables, parameters, or both. One possibility for computing the Bayes factor between model A and model B is to average the quantities $m(\mathbf{y}, \mathbf{z}^b|M_A)/m(\mathbf{y}, \mathbf{z}^b|M_B)$ for each iteration b, where $\mathbf{z}^b$ is the b^{th} posterior draw of z from model B (Raftery 1993). It is often simple to construct these quantities with a Laplace approximation or through other meth-

ods (Kass and Raftery 1995, p. 780). However, this method might be ineffective if approximation of $m(\mathbf{y}, \mathbf{z}^b|M_A)/m(\mathbf{y}, \mathbf{z}^b|M_B)$ for each b is not very fast, owing to the large number of approximations that must be performed. It is also possible that the Laplace approximation might not be very accurate either.

Another alternative for estimation of marginal likelihoods was proposed by Chib (1995). This approach is based on a clever use of Bayes' Theorem. If the likelihood function of data $\mathbf{y}$ depends on a parameter vector $\boldsymbol{\theta}$, then Bayes Theorem establishes that $\pi(\boldsymbol{\theta}|\mathbf{y}) = p(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})/m(\mathbf{y})$, and solving for the marginal likelihood $m(\mathbf{y})$ is trivial. After taking the natural logarithm of both sides of the equality, it is evident that $\log(m(y)) = \log(p(\mathbf{y}|\boldsymbol{\theta}) + \log(\pi(\boldsymbol{\theta})) - \log(\pi(\boldsymbol{\theta}|\mathbf{y}))$. The truthfulness of this identity holds for any $\boldsymbol{\theta}_0$ that is in the support of the posterior (and thus naturally the prior as well). After obtaining this formulation for $\log(m(y))$, Chib proposed estimation of the log marginal likelihood by estimating/evaluating each of the three summands separately and then adding them. It is often straightforward to evaluate the prior density and the likelihood for a particular value of $\boldsymbol{\theta}_0$. An estimate of the posterior density at $\boldsymbol{\theta}_0$ is gotten, and then each of the quantities needed to estimate the log marginal likelihood is in place.

Chib (1995) discussed approaches to estimation of the posterior at a given point when all full conditionals are known. Later, Chib and Jeliazkov (2001) extended the approach to the case where full conditionals are known only up to a normalizing constant factor. These approaches can accommodate latent variables. All latent variables might be integrated out when evaluating the likelihood (Chib and Jeliazkov 2001, p. 272). Alternatively, some of the latent variables may be combined with $\boldsymbol{\theta}_0$ if it is too difficult to evaluate the likelihood otherwise (Chib 1995, p. 1316). However, neither of these approaches is satisfactory in dealing with rank data because (1) there are at least as many latent variables as observations and (2) they are not able to be

integrated out in closed form except perhaps in overly simplistic models that do not allow ties.

In situations where approximation of the Bayes factor is not practical or not reliable, an alternative for model comparison between nested models is to use discrepancy measures based on pivotal quantities. Johnson (2007) demonstrated how useful discrepancy measures based on pivotal quantities can be for assessing model assumptions. The basic idea starts with choosing a pivotal quantity that allows model assumptions to be checked with quantities that have a known distribution when the model is correct. Johnson proved, under some mild conditions, that if the model is correct then the pivotal quantity will have an identical distribution whether it is computed using the true model parameters or a posterior draw (pp. 720–721). This result is very powerful because while the true parameters are unknown, posterior draws might be obtained using direct sampling or an MCMC algorithm.

For example, if it is assumed that $\mathbf{y}|\{\mathbf{X}, \boldsymbol{\beta}, \sigma^2\} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, then the vector of standardized residuals $\mathbf{r} \equiv (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})/\sigma$ is assumed to follow the $\sim N(\mathbf{0}, \mathbf{I})$ distribution (p. 728). The known joint distribution of the standardized residuals does not depend on any model parameters and thus any function of these standardized residuals will be a pivotal quantity if the model is correct.

In a typical Bayesian Monte Carlo approach to posterior inference, many draws from the posterior distribution are required. It might seem disconcerting, then, that there are possibly as many unique values of the pivotal quantity corresponding to these draws as there are draws themselves! Rather than arbitrarily choosing one of them, all of them should be used. This is delicate because, as Johnson noted, the pivotal quantities are interrelated in that they are all based on the same observed data. Nonetheless, Johnson applied results from other authors to obtain an upper bound on the prior-predictive-posterior (PPP) p-value, a p-value that can be used

in assessing Bayesian models. Let $F^{-1}(\cdot)$ be the quantile function associated with a pivotal quantity where larger values are expected when the model is wrong. Out of n evaluations of the pivotal quantity (at each of n posterior draws), let $PQ_{(m),n}$ represent the m^{th} largest. For $u \in (0,1)$,

$$Pr(PQ_{(m),n} > F^{-1}(u)) \leq \min(1, (n - nu)/(n - m + 1)), \tag{2.4}$$

and so (for u not too close to 1) if n is large, m/n is bounded under 1, $PQ_{(m),n} > F^{-1}(u)$, and $(1 - m/n) >> (1 - u)$, there is an indication that the model is flawed (p. 724).

D. Non-local Priors

A key contribution of the proposed work, in addition to the methodological innovation of jointly modeling binomial and rank data, is the parsimonious use of non-local priors on the variance parameters of random effects. Johnson and Rossell (2010) define local and non-local alternative priors in a Bayesian hypothesis test. Consider testing for the presence of a random effect due to the P^{th} factor, where σ_P^2 denotes the variance of this random effect. Under the simpler model, say M_0, the prior distribution for σ_P^2 is degenerate at 0. A local alternative prior for M_1 would have a prior distribution that is positive at the point $\sigma_P^2 = 0$. A non-local alternative prior for M_1 would vanish to (equal) 0 as σ_P^2 approaches (equals) 0. A compelling rationale for employing non-local priors in hypothesis testing is that they can be used to prevent inordinate imbalances in the rate of accumulation of evidence for or against a simpler model (Johnson and Rossell 2010, p. 148).

One difficulty with using non-local priors is that they might preclude the use of certain techniques for calculating Bayes factors. For example, Verdinelli and Wasser-

man (1995) proposed estimating the Bayes factor using the product of estimates of the Savage-Dickey density ratio and a correction term, but their approach requires that the parameter space for one model be nested in that of the other (p. 618). Meng and Wong (1996) introduced the technique of bridge sampling to estimate Bayes factors; the technique is characterized by sampling from each of the two models considered as well as a collection of intermediate models. However, this approach is not advised for the proposed methodology because it is suited for comparing models with overlapping parameter spaces (e.g., local priors). By construction, non-local priors use mutually exclusive parameter spaces. While some Bayes factor estimating techniques are therefore precluded, discrepancy measures may still be used because the proposed methodology compares models which may be seen as reductions or extensions of other models although they are not strictly nested. Discrepancy measures provide a viable alternative as they can examine models for the need of a particular extension.

E. Primate Intelligence

Aside from the more statistically oriented discussion of the previous sections, some brief comments about the study of primate intelligence are in order because of the motivating application for the present work.

Spearman (1904) combined information from various assessments to understand "general intelligence" in humans. This work is credited with introducing the widely used statistical technique of factor analysis. A tremendous amount of research has been devoted to studying intelligence since then, not to mention the research preceding and concurrent to Spearman's seminal work.

The study of intelligence has also been applied to other species, and in particular we note such studies in non-human primates. Among the possible purposes of animal

research are comparisons between primate orders (Harlow and Mears 1979, p. 1). The great apes and monkeys have been studied using a variety of assessments, or *procedures*, which correspond to different *paradigms* (Johnson et al. 2002) of what we loosely refer to as *intelligence* but might properly be termed *cognition, learning, inhibition, memory,* or something else entirely, depending on the particular paradigm.

Primate intelligence studies have not only been used to make comparisons between species but have also tried to determine plausible explanations for them. For example, Amici et al. (2008) found an association between inhibitory ability and fission-fusion prevalence—the degree to which primates form and disband in different group structures.

Still more work addresses the question of whether there is one general intelligence construct or paradigm-specific intelligences. For example, Riopelle and Hill (1973) claimed that intelligence cannot be represented by one trait in humans nor in animals (p. 541). However, Johnson et al. (2002) found little to no evidence of significant genus-paradigm interactions in their meta-analysis of non-human primate intelligence data.

CHAPTER III

METHODOLOGY

This chapter presents the proposed methodology for joint modeling of binomial and rank data. Section A presents the likelihood and prior distributions for the model allowing an arbitary number of random effects. Section B presents details of a Metropolis-within-Gibbs MCMC algorithm that may be used for inference on the posterior distribution. Section C presents an extension of the model to allow for assessment-specific conditional error variances, including the minor differences in the prior distribution and the MCMC algorithm. Section D discusses model comparison, giving special attention to the use of discrepancy measures based on pivotal quantities. The proposed methodology is applied in Chapter IV to primate intelligence data.

A. Model Specification

The modeling entails a substantial amount of notation, which is now presented.

Suppose that data come 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 $T_j \equiv 1$. 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 only, the t subscript is sometimes omitted for simplification (e.g., y_{ij} instead of y_{ijt}). Rank data are assumed to be ordered so that higher values of y_{ij} reflect better—not worse—performance.

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 with nonmissing response for that assessment.

1. Likelihood

A key feature in joint modeling of binomial and rank outcomes is the use of latent variables to facilitate modeling. As discussed in Chapter II, it is possible to model binomial data using multiple latent variables, and if the latent variables have a conditionally normal distribution, then posterior inference on parameters affecting the probability of success can be quite convenient. Similarly, rank data can be modeled using latent variables with conditionally normal distributions that conveniently allow the probability of any given ranking to depend on parameters.

Recall the definition of the quantity $p_{(i)}(\kappa)$ in Equation 2.1. This is a simplification of the the quantity $p_{(i,j)}(\kappa)$ that was originally proposed by Johnson et al. (2002, eq. 2) to allow for multiple assessments with the subscript j. If $z_{(i),j}$ ($y_{(i),j}$) is the i^{th} largest latent (observed) variable for rank response assessment j, then

$$p_{(ij)}(\kappa_j) = \begin{cases} \exp(-(z_{(i+1),j} - z_{(i),j})/\kappa_j) & \text{if } y_{(i+1),j} = y_{(i),j} \\ 1 - \exp(-(z_{(i+1),j} - z_{(i),j})/\kappa_j) & \text{if } y_{(i+1),j} > y_{(i),j}. \end{cases} \tag{3.1}$$

Equation 3.1 differs from the original formulation by Johnson et al. (2002) in that it uses an assessment-specific value of κ and therefore is appropriate when the proportion of ties is quite different across assessments. In a multiple-assessment situation, rank data might be missing for some assessments; the $z_{(i),j}$'s and $y_{(i),j}$'s use only the nonmissing y_{ij}'s and the corresponding z_{ij}'s.

For both binomial and rank data, the likelihood for the observed responses y_{ijt} depends on the associated z_{ijt}. In addition, the likelihood depends upon the assessment-

specific cutpoint parameters τ_j for binomial responses and the assessment-specific parameters κ_j that influence the probability of ties for rank responses. To accommodate missing responses, let $w_{ijt} = 1$ if y_{ijt} is observed and 0 otherwise. Note that the likelihood, defined in Equation 3.2, is not affected by any z_{ijt} values for which $w_{ijt} = 0$.

$$f(\mathbf{y}|\mathbf{z}, \boldsymbol{\tau}, \boldsymbol{\kappa}) = \left[\prod_{j:B(j)=1} \prod_{i=1}^{I} \prod_{t=1}^{T_j} (1(\{y_{ijt} = 0 \cap z_{ijt} \leq \tau_j\} \cup \{y_{ijt} = 1 \cap z_{ijt} > \tau_j\}))^{w_{ijt}} \right] \times \left[\prod_{j:B(j)=0} \left(\prod_{i=1}^{C(j)-1} p_{(ij)}(\kappa_j) \right) \left(\prod_{i=1}^{I} \prod_{i':z_{i'j}<z_{ij}} (1(y_{i'j} \leq y_{ij}))^{w_{ij}w_{i'j}} \right) \right] \tag{3.2}$$

The first part of the likelihood is for binomial data (see Albert and Chib 1993) and the second is for rank data (see Johnson et al. 2002, eq. 3). The ability of the joint likelihood to handle missing data that are assumed to be missing completely at random stems from the ability to handle them in each of the response types using existing techniques. 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 $\mathbf{z}$.

2. Hierarchical Model Priors

A principal advantage afforded by the latent variables z_{ijt} is that they may be modeled in the same manner regardless of whether they correspond to rank or binomial data. Higher values of y_{ijt} are better for each data type, and the form of the likelihood ensures that the same is true of the latent data. Thus, a data set of mixed response type can seamlessly be used to assess questions of primary interest, such as whether a random effect influences the latent variables and therefore the responses.

The latent variables must have their location and scale established through informative priors, if not exact constraints. It is well known that the likelihood of binomial data, as in the first part of Equation 3.2, is invariant to (a) adding a constant to each z_{ijt} and to each τ_j, and (b) multiplying each z_{ijt} and each τ_j by any positive constant. In addition, the likelihood of the rank data, as in the latter part of Equation 3.2, is also invariant to (a) a location transformation of each z_{ijt} and (b) a scale transformation of each z_{ijt} and each κ_j.

Although the location and scale of the latent variables is not inherently established, various possibilities exist to ensure identifiability of the model parameters. Among these many options, it is preferable that each latent variable have the same marginal mean and marginal variance under the prior distribution to aid in interpretability and comparison of model parameters. Throughout this dissertation, unless stated otherwise the marginal mean and marginal variance refer to the mean and variance of the marginal distribution of z_{ijt} under the *prior*. Among the simplest of models for $\mathbf{z}$ satisfying the condition of common marginal means and marginal variances, suppose

$$\pi(\mathbf{z}) = \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{t=1}^{T_j} (2\pi)^{-1/2} \exp(-z_{ijt}^2/2). \tag{3.3}$$

In Equation 3.3, the prior implies that each latent variable is *a priori* independent with a standard normal distribution. Of course, if this prior is used it is advisable that the priors on each κ_j and τ_j not be too peaked as the model would otherwise be far too rigid. In particular, the common practice of constraining $\tau_j \equiv 0$ would be inappropriate unless it is desired to assume that the posterior predictive probability of a success must be exactly one-half, hardly a tenable assumption.

It is possible to extend the model for $\mathbf{z}$ in Equation 3.3 while still allowing

each latent variable to have the same marginal distribution. This can be accomplished through hierarchical modeling. Consider first a simple extension involving one subject-specific random effect, $u_{1,i}$. Assume

$$z_{ijt} = u_{1,i} + \epsilon_{ijt}$$

with $\mathbf{u}_1 \sim N(\mathbf{0}, \sigma^2_{u,1}\mathbf{I})$, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2_\epsilon \mathbf{I})$, and $\mathbf{u}_1$ and $\boldsymbol{\epsilon}$ conditionally mutually independent. Conditional on $\mathbf{u}_1$, $\sigma^2_{u,1}$, and σ^2_ϵ,

$$\pi(\mathbf{z}|\mathbf{u}_1, \sigma^2_{u,1}, \sigma^2_\epsilon) = \prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{t=1}^{T_j} (2\pi\sigma^2_\epsilon)^{-1/2} \exp(-(z_{ijt} - u_{1,i})^2/2\sigma^2_\epsilon).$$

Marginalizing over the random effects $\mathbf{u}_1$, it is easily seen that $\mathbf{z}|\{\sigma^2_{u,1}, \sigma^2_\epsilon\} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where each diagonal element of $\boldsymbol{\Sigma}$ equals $\sigma^2_{u,1}+\sigma^2_\epsilon$. Requiring that $\sigma^2_{u,1}+\sigma^2_\epsilon = 1$ ensures that each z_{ijt} has the same marginal distribution (standard normal).

In the same manner, it is possible to include an arbitrary number of random factor effects. Suppose that there are P random effects, and the p^{th} random effect has L_p levels. An explicit notation might use $u_{p,l(p,i,j)}$ to represent the p^{th} random effect at level l, which level l might depend on the random effect, subject, assessment, or some combination thereof. The random effects model for the latent variables z_{ijt} with P random effects is straightforward.

$$z_{ijt} = \sum_{p=1}^{P} u_{p,l(p,i,j)} + \epsilon_{ijt} \tag{3.4}$$

Assume that $\boldsymbol{\epsilon}|\sigma^2_\epsilon \sim N(\mathbf{0}, \sigma^2_\epsilon \mathbf{I})$, that $\mathbf{u}_p|\sigma^2_{u,p} \sim N(\mathbf{0}, \sigma^2_{u,p}\mathbf{I})$ for $p = 1, 2, \ldots, P$, and that all random effects and error terms are conditionally mutually independent. These conditions, together with Equation 3.4, define the distribution of each latent variable

conditional on the variance parameters but marginalized over the random effects.

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

The constraint $\sigma^2_\epsilon + \sum_p \sigma^2_{u,p} = 1$ ensures that the scale is established, with all latent variables having the same marginal expectation and the same marginal variance.

We emphasize that the common marginal mean and marginal variance across models is advantageous, particularly with the constraint that the marginal variance equal the arbitrary value of 1. Each random effect's variance parameter can then be interpreted as the prior proportion of total variability in $\mathbf{z}$ that is attributed to that random effect. Also, by fixing the scale across models with differing numbers of random effects, it is more sensible to choose a prior density for each κ_j and each τ_j that does not depend on the model inasmuch as these parameters are tied to the scale of $\mathbf{z}$. Comparison of elements in $\mathbf{z}$ is likewise aided by the common marginal mean and marginal variance as they are readily comparable across models and across assessments.

Unfortunately, the introduction of fixed effects (i.e., $\mathbf{X}\boldsymbol{\beta}$) or random coefficients (i.e., $W\mathbf{u}$ where W contains continuously measured values) can considerably impact the proposed identification of location and scale. Suppose that $Z|\mathbf{X}, \boldsymbol{\beta}, \Sigma \sim N(X\beta, \Sigma)$ and $\beta|\Sigma \sim N(\mu_\beta, S_\beta)$. Then $\mathbf{z}|\Sigma, \mu_\beta, S_\beta \sim N(X\mu_\beta, XS_\beta X' + \Sigma)$. In general, the latent variables might have different marginal scales—particularly if X contains continuously measured covariates—and might have different marginal means. Different marginal scales will generally result when random coefficients models are used, as well. When working with binomial data in a fixed effects model, it is common to induce identifiability of the latent variables by assuming that $diag(\mathbf{\Sigma}) = \mathbf{1}$ and $\tau_j = 0$ for each assessment, while sometimes also assuming the prior on β is uniform (e.g., Albert

and Chib 1993, p. 671). However, the proposed methodology considers only models involving random factor effects because such models can easily enforce a common scale in the latent variables. As mentioned earlier, the advantages of the common scale are manifest in elication of the prior distribution, comparison of models, and interpretation of random effects.

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

$$\pi(\mathbf{z}, \mathbf{u}_1, \ldots, \mathbf{u}_P, \sigma^2_{u,1}, \ldots, \sigma^2_{u,P}, \sigma^2_\epsilon, \boldsymbol{\tau}, \boldsymbol{\kappa}). \tag{3.5}$$

The remainder of the prior specification will assume that Equation 3.5 may be factored as

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

As implied by Equation 3.4 and the surrounding text,

$$\pi(\mathbf{z}|\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_P, \sigma^2_\epsilon) = \prod_{i=1}^{I}\prod_{j=1}^{J}\prod_{t=1}^{T_j} (2\pi\sigma^2_\epsilon)^{-1/2} \exp(-(z_{ijt} - \sum_p u_{p,l(p,i,j)})^2/2\sigma^2_\epsilon) \tag{3.6}$$

and

$$\pi(\mathbf{u}_p|\sigma^2_{u,p}) = \prod_{l=1}^{L_p} (2\pi\sigma^2_{u,p})^{-1/2} \exp\left(-u^2_{1,p}/2\sigma^2_{u,p}\right), \quad \forall p = 1, \ldots, P. \tag{3.7}$$

These priors for the latent variables and random effects are particularly convenient for use in posterior sampling, as is further explained in Section B.

A Dirichlet prior is chosen for the prior distribution of the variance parameters. By so doing, the constraint that the variances sum to 1 is enforced. Furthermore, the Dirichlet density can readily be used as a non-local prior. Let $\boldsymbol{\sigma} \equiv (\sigma^2_\epsilon, \sigma^2_{u,1}, \ldots, \sigma^2_{u,P})$ 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

$$\pi(\boldsymbol{\sigma}) = \frac{\Gamma(\alpha_\epsilon + \alpha_{u,1} + \cdots + \alpha_{u,P})}{\Gamma(\alpha_\epsilon)\Gamma(\alpha_{u,1})\cdots\Gamma(\alpha_{u,P})}(\sigma_\epsilon^2)^{\alpha_\epsilon - 1}(\sigma_{u,1}^2)^{\alpha_{u,1}-1}\cdots(\sigma_{u,P}^2)^{\alpha_{u,P}-1}. \tag{3.8}$$

Inspection of this density reveals that for any variance with associated parameter $\alpha > 1$, the prior density approaches (equals) 0 as the variance approaches (equals) 0, thus demonstrating its usefulness as a non-local prior. Testing the need for a particular random effect is equivalent to testing the hypothesis that the variance parameter for that random effect is positive. By using a non-local prior on $\boldsymbol{\sigma}$, a correct null hypothesis that a random effect is not present is better able to gain support from the data.

While generally the priors to this point have been chosen because of their convenience in interpretation or in posterior sampling, no obviously advantageous priors are apparent for either the κ_j's nor the τ_j's. Let $\boldsymbol{\kappa}$ be the collection of κ_j's for all assessments j with rank response, and let $\boldsymbol{\tau}$ be the collection of τ_j's for all assessments j with binomial response. A reasonable choice for a prior on $\boldsymbol{\kappa}$ is to assume that the κ_j are mutually independent with a $Gamma(a_j, b_j)$ distribution. The multivariate gamma prior ensures that none of the κ_j values are negative. If the κ_j are considered exchangeable, a_j and b_j could be replaced by a and b.

$$\pi(\boldsymbol{\kappa}) = \prod_{j:B(j)=0} {b_j}^{a_j}(\Gamma(a_j))^{-1}{\kappa_j}^{a_j-1}\exp(-b_j\kappa_j)1(\kappa_j > 0) \tag{3.9}$$

A sensible prior for $\boldsymbol{\tau}$ is to assume that each is *a priori* independent and has a $Cauchy(m_j, s_j^2)$ distribution. The relatively thick tails characteristic of the Cauchy distribution are desirable if the prior is intended to be somewhat informative but not overly restrictive. If the τ_j are considered exchangeable, m_j and s_j^2 should have

common values across assessments.

$$\pi(\boldsymbol{\tau}) = \prod_{j:B(j)=1} (2\pi(s_j^2)^{1/2}(1 + (\tau_j - m_j)^2/s_j^2))^{-1} \tag{3.10}$$

The form of the joint model for binomial and rank data is now completely specified. All that remains is determining which random effects will be used and selecting values for all hyperparameters—the α values in the prior for $\boldsymbol{\sigma}$, the a_j and b_j values in the prior for $\boldsymbol{\kappa}$, and the m_j and s_j^2 values in the prior for $\boldsymbol{\tau}$. Selection of the hyperparameters requires care, and ideally is application-specific and based on subject knowledge. Model selection strategies can be considered to evaluate which random effects should be included in the model. For example, to perform a Bayesian hypothesis test that a random effect is needed in the model, the Bayes factor might be computed to compare the models with and without the random effect. As stated earlier, a non-local prior should be used on the variance of the questioned random effect, which is achieved by selecting the corresponding α to be strictly greater than 1. Discrepancy measures based on pivotal quantities are recommended if computation of Bayes factors proves difficult, which is likely to be the case in large samples.

Hyperparameter and model selection is described in Chapter IV for analysis of primate intelligence data.

B. Posterior Inference Using Markov Chain Monte Carlo

In the Bayesian paradigm, inference on model parameters is made using the posterior distribution. For this model, the posterior distribution (up to a normalizing constant) is given by Equation 3.11.

$$p(\mathbf{y}|\mathbf{z}, \boldsymbol{\kappa}, \boldsymbol{\tau})\pi(\mathbf{z}|\mathbf{u}_1, \ldots, \mathbf{u}_P, \sigma_\epsilon^2)\pi(\sigma_\epsilon^2, \sigma_{u,1}^2, \ldots, \sigma_{u,P}^2)\pi(\boldsymbol{\kappa})\pi(\boldsymbol{\tau}) \prod_{p=1}^{P} \pi(\mathbf{u}_p|\sigma_{u,p}^2) \tag{3.11}$$

Table I. 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

This model is too complex to analytically determine the exact posterior distribution of the parameters and latent variables because of the unknown normalizing constant implicit in Equation 3.11. As a result, model inferences could be based on samples from the posterior distribution. An exact sampling technique is not readily available, but a Markov chain Monte Carlo (MCMC) algorithm can be implemented to produce a series of correlated draws with limiting distribution equal to the exact posterior.

The algorithm employed uses a Metropolis-within-Gibbs sampling technique. With one exception, the algorithm updates parameters and latent variables by sampling from full conditional distributions exactly or with the Metropolis-Hastings algorithm. The choice of priors mentioned in Section A was made with an eye towards making updates easy. An overview of the MCMC algorithm appears in Table I, while the details are contained in the remainder of this section.

In each of the updating steps, the algorithm uses the current values of various parameters/latent variables. Any quantity that is conditioned on is implicitly using the most recent value of that quantity unless stated otherwise. When it is necessary to

distinguish between the current value and the proposed value of an arbitrary quantity ξ, ξ^c will represent the current value and ξ^* will represent the proposed value. When it is important to identify the iteration number, $\xi^{(m)}$ represents the value of ξ at the end of the m^{th} iteration. At the end of each iteration, $\xi^{(m)}$ must equal ξ^c because ξ^c is updated whenever ξ changes.

The iterative MCMC algorithm requires that the chain first be initialized. The level of sophistication used to do so may vary, but a straightforward (if inefficient) initialization might be done as follows:

1. Set $\tau_j^{(0)} = m_j$ for each $j : B(j) = 1$.
2. Set $\kappa_j^{(0)} = a_j/b_j$ for each $j : B(j) = 0$.
3. Set $\mathbf{u}_1^{(0)} = \mathbf{0}, \ldots, \mathbf{u}_P^{(0)} = \mathbf{0}$.
4. Set $(\sigma_\epsilon^2)^{(0)} = (\sigma_{u,1}^2)^{(0)} = \cdots = (\sigma_{u,P}^2)^{(0)} = 1/(P+1)$.
5. Set each $z_{ijt}^{(0)}$.
    - For assessments such that $B(j) = 1$, set $z_{ijt}^{(0)} = \tau_j^{(0)} + 0.5$ if $y_{ijt} = 1$, set $z_{ijt}^{(0)} = \tau_j^{(0)} - 0.5$ if $y_{ijt} = 0$, and set $z_{ijt}^{(0)}$=0 if y_{ijt} was not observed.
    - For assessments such that $B(j) = 0$, set $z_{ijt}^{(0)} = [2y_{ijt} - (1 + C(j))]/C(j)$. Because the methodology assumes that rank data is ordered from worst ($y_{ijt} = 1$) to best, the initialized values will be between -1 and 1 and obey the posterior ordering constraints. If y_{ijt} was not observed, set $z_{ijt}^{(0)} = 0$.

This initialization of the latent variables is underdispersed because an overdispersed initialization might get stuck. Certainly the overall initialization might be refined, but the preceding can serve as a default method.

Different techniques are used to update $\mathbf{z}$, depending on whether the latent variables correspond to binomial or rank data. First, consider binomial data. There is an intricate relationship between the cutpoint parameters $\boldsymbol{\tau}$ and the latent variables $\mathbf{z}$. It is possible to update each cutpoint parameter using its full conditional (a truncated Cauchy). Then the latent variables for binomial data can be updated using their full conditionals (truncated normals). These full conditionals are now stated; for convenience, parameters and latent variables not appearing in the functional form of the full conditionals are not explicitly stated.

$$\pi(\tau_j|\mathbf{z},\mathbf{y}) \propto Cauchy(m_j, s_j^2)1(\tau_j \in (l_j^c, u_j^c)), \quad \forall j : B(j) = 1 \tag{3.12}$$

$$\begin{aligned}\pi(z_{ijt}|\tau_j, \mathbf{u}_1, \ldots, \mathbf{u}_P, \sigma_\epsilon^2, \mathbf{y}) \propto N(u_{1,l(1,i,j,t)} + \cdots + u_{P,l(P,i,j,t)}, \sigma_\epsilon^2) \\ \times 1(z_{ijt} \in (L_{ijt}^c, U_{ijt}^c)) \quad \forall i,t,j : B(j) = 1\end{aligned} \tag{3.13}$$

The truncation regions are dependent on the current values of $\mathbf{z}$ or $\boldsymbol{\tau}$.

$$l_j^c = \max(-\infty, \max_{i,t}\{z_{ijt}^c : y_{ijt} = 0\})$$

$$u_j^c = \min(\infty, \min_{i,t}\{z_{ijt}^c : y_{ijt} = 1\})$$

$$L_{ijt}^c = \begin{cases} -\infty & \text{if } y_{ijt} = 0 \text{ or is missing} \\ \tau_j^c & \text{if } y_{ijt} = 1 \end{cases}$$

$$U_{ijt}^c = \begin{cases} \tau_j^c & \text{if } y_{ijt} = 0 \\ \infty & \text{if } y_{ijt} = 1 \text{ or is missing} \end{cases}$$

It is conceptionally easy to update these cutpoint parameters and latent variables. However, the approach of updating $\boldsymbol{\tau}$ and $\mathbf{z}$ with Gibbs sampling in separate blocks is problematic because it tends to mix poorly (see Cowles 1996, p. 104).

Such recognized inefficiency led Cowles (1996) to explore an alternative strategy

for MCMC sampling of ordinal probit models, which is equally applicable to the special case of binomial data when the cutpoint is not fixed. Cowles proposed that instead of updating the cutpoint parameters and the latent variables in separate blocks, they be updated in one block. However, instead of simultaneously sampling both $\mathbf{z}$ and $\boldsymbol{\tau}$ from $\pi(\mathbf{z}, \boldsymbol{\tau}|$ all else), this step was split into sampling from $\pi(\boldsymbol{\tau}|$all else except $\mathbf{z})$ and then from $\pi(\mathbf{z}|$ all else). The implication is that by marginalizing over $\mathbf{z}$ in the full conditional of $\boldsymbol{\tau}$, exact sampling is no longer practical. Cowles used a Metropolis-Hastings step to simultaneously update the multiple cutpoints in her ordinal probit setting, and then updated the latent variables only if the proposed cutpoint draws were accepted.

Unlike the scheme proposed by Cowles (1996), the following sampling scheme updates the latent variables regardless of whether or not the proposed cutpoint draws were accepted and thus matches the implementation of Cowles' algorithm explained by (Johnson and Albert 1999, pp. 135–136).

1. Update $\boldsymbol{\tau}$ by individually updating each τ_j with a Metropolis step using a normal distribution as the proposal distribution.

2. Regardless of whether $\boldsymbol{\tau_j}$ was accepted in the previous step, update each z_{ijt} : $B(j) = 1$ using its complete conditional.

A potential advantage of updating z_{ijt} in every instance is better chain mixing. Although this mandatory update is not required for the algorithm to be valid, the consequences of less frequent updating of the latent variables might involve poorer mixing of other model unknowns, such as the random effects.

The details for updating each τ_j with a Metropolis step are adapted from Cowles' approach. Let $u_{ijt} \equiv \sum_p u_{p,l(p,i,j)}$ be the sum of the random effects, and let $f(\tau_j)$ be

the unnormalized full conditional of τ_j after integrating over the latent variables $\mathbf{z}$.

$$f(\tau_j) = (1 + (\tau_j - m_j)^2/s_j^2)^{-1} \times \prod_{i=1}^{I} \prod_{t=1}^{T_j} \left[\left(\Phi\left((\tau_j - u_{ijt})/\sigma_\epsilon \right) \right)^{1-y_{ijt}} \left(1 - \Phi\left((\tau_j - u_{ijt})/\sigma_\epsilon \right) \right)^{y_{ijt}} \right]^{w_{ijt}} \tag{3.14}$$

As usual, $\Phi(\cdot)$ represents the cdf of the standard normal. To set $\tau_j^{(m)}$, the proposed value τ_j^* is sampled from the $N(\tau_j^c, v_j^2)$ distribution, where v_j^2 is a tuning parameter. Note that this proposal distribution is symmetric; the transition kernel q has the property that $q(\tau_j^*|\tau_j^c) = q(\tau_j^c|\tau_j^*)$. Thus, the Metropolis ratio implies that the acceptance probability is $\min(1, f(\tau_j^*)/f(\tau_j^c))$. With this probability, set $\tau_j^{(m)} = \tau_j^*$, and otherwise set $\tau_j^{(m)} = \tau_j^c$.

After updating $\boldsymbol{\tau}$, each z_{ijt} for binomial data is updated using its full conditional (Equation 3.13). This is conceptionally straightforward because the full conditional is in each case a truncated normal distribution. Actual implementation might use rejection sampling or evaluation of the quantile function at a randomly chosen point.

- Sample from the $N(u_{ijt} = u_{1,l(1,i,j)} + \cdots + u_{P,l(P,i,j)}, \sigma_\epsilon^2)$ distribution until a draw is found in (L_{ijt}^c, U_{ijt}^c). For regions where draws are rarely accepted consider the algorithm of Robert (1995), which uses rejection sampling with draws generated from an exponential distribution.

- Let $F(\cdot)$ be the cumulative distribution function and $F^{-1}(\cdot)$ the quantile function of the $N(u_{ijt}, \sigma_\epsilon^2)$ distribution. Use $F^{-1}(s)$, where s is sampled from the $Uniform(F(l_{ijt}^c), F(u_{ijt}^c))$ distribution.

An updating procedure for $\boldsymbol{\tau}$ and the latent variables for binomial data has been detailed; we now demonstrate how the latent variables for rank data may be updated. The first part of the procedure is given by Johnson et al. (2002). A Metropolis-

Hastings update is used, with each $z_{ijt} : B(j) = 0$ being individually updated. Recall that the t is optional for rank data because then $t \equiv 1$. The proposal distribution for z_{ij} is the $N(u_{1,l(1,i,j)} + \cdots + u_{P,l(P,i,j)}, \sigma_\epsilon^2)$ distribution, truncated to the region

$$\left(\max(-\infty, \max_{i':y_{i'j}<y_{ij}} z_{i'j}), \min(\infty, \min_{i':y_{i'j}>y_{ij}} z_{i'j})\right). \tag{3.15}$$

If y_{ij} is missing, z_{ij} does not affect the proposal distribution's truncation region for any of the latent variables, and also the truncation region for updating z_{ij} is defined as $(-\infty, \infty)$. To calculate the acceptance probability, let $\mathbf{z}^c$ be the collection of current values of z_{ijt}, and let $\mathbf{z}^*$ be the collection of candidate (proposed) values. Note that because the z_{ijt} are individually updated, $\mathbf{z}^c$ and $\mathbf{z}^*$ will differ by at most a single element. Furthermore, upon updating each z_{ij}, both $\mathbf{z}^c$ and $\mathbf{z}^*$ are also updated. Let $p_{(ij)}(\kappa_j)^c$ be $p_{(ij)}(\kappa_j)$ when using the values of κ_j^c and $\mathbf{z}^c$, and let $p_{(ij)}(\kappa_j)^*$ be $p_{(ij)}(\kappa_j)$ when using the values of κ_j^c and $\mathbf{z}^*$. The acceptance probability for $z_{ij}^* : B(j) = 0$ is

$$\min\left(1, \left[\prod_{i=1}^{C(j)} p_{(ij)}(\kappa_j)^*\right]\left[\prod_{i=1}^{C(j)} p_{(ij)}(\kappa_j)^c\right]^{-1}\right). \tag{3.16}$$

If y_{ij} is missing then the acceptance probability for z_{ij} is always one.

Because latent variables are individually updated, it might be difficult for a group of latent variables with the same observed response to effectively traverse the support. For example, consider the possibility that two observations, say y_{1j} and y_{2j}, are tied for being the worst in assessment j, but the proportion of ties is low because of a very small value of κ_j. In an overdispersed initialization, it is possible that $z_{(1)j}^{(0)}$ and $z_{(2)j}^{(0)}$ are both much lower than $z_{(3)j}^{(0)}$. Suppose they are -5.1, -5.09, and -1.8. If the proposal distribution for $z_{1j}^{(1)}$ is very concentrated around, say, -2.0, the proposed value might be unlikely to be accepted because a value near -2.0 would make $p_{(1j)}(\kappa_j^{(0)})$ very small. But likewise, if the proposal distribution for $z_{2j}^{(1)}$ is concentrated around

-2.0, a proposed value near -2.0 would again cause $p_{(1j)}(\kappa_j^{(0)})$ to be very small. The conundrum, then, is that the proposal distributions might favor values of z_{1j} that are far from z_{2j} (and vice versa) but the likelihood might be dramatically smaller if z_{1j} and z_{2j} are not close. This would make it difficult for either z_{1j} or z_{2j} to move. This problem can persist throughout any finite run of the MCMC algorithm.

To circumvent this difficulty, we add an extra step to the procedure given by Johnson et al. (2002) for updating the z_{ij}'s associated with rank data. This step allows shifts in z_{ij} values from assessment j that have the same y_{ij} values. The extra step is not essential for the algorithm to be valid and so it may be omitted. However, it is recommended to help with chain mixing. Let y_j be a unique observed value of the assessment j responses. After updating each individual z_{ij} for which $y_{ij} = y_j$ (and thus each such z_{ij}^c), an additive shift of δ is proposed for the collection of such z_{ij}'s.

$$\forall i, \quad z_{ij}^* = \begin{cases} z_{ij}^c + \delta, \text{ if } y_{ij} = y_j \\ \\ z_{ij}^c, \text{ otherwise} \end{cases}$$

Care is taken in choosing δ's proposal distribution to prevent a proposed shift that is inconsistent with the observed rankings. The proposal distribution is a normal distribution with mean 0 and variance given by tuning parameter v^2, truncated to the region that ensures appropriate ordering on the latent variables. The lower and upper limits, given by Equation 3.17 and Equation 3.18, are denoted by LL_{yj}^c and UL_{yj}^c to emphasize that they are dependent on the current values of all z_{ij}'s with observed rankings below or above the unique y_j value being considered.

$$LL_{yj}^c = \left(\max_{i':y_{i'j}<y_j} z_{i'j}^c - \min_{i':y_{i'j}=y_j} z_{i'j}^c \right) \tag{3.17}$$

$$UL_{yj}^c = \left(\min_{i':y_{i'j}>y_j} z_{i'j}^c - \max_{i':y_{i'j}=y_j} z_{i'j}^c \right) \tag{3.18}$$

The Metropolis-Hastings ratio used in the acceptance probability also depends on LL^*_{yj} and UL^*_{yj}, which are analagously defined. The acceptance probability for the collection $\{z^*_{ij} : y_{ij} = y_j\}$ is the minimum of one and the result of Equation 3.19.

$$\frac{\left(\prod_{i=1}^{C(j)-1} p_{(ij)}(\kappa_j)^*\right)\left(\prod_{i=1}^{I}\exp(-(z^*_{ij}-u_{1,l(1,i,j)}-\cdots-u_{P,l(P,i,j)})^2/2\sigma^2_\epsilon)\right)}{\left(\prod_{i=1}^{C(j)-1} p_{(ij)}(\kappa_j)^c\right)\left(\prod_{i=1}^{I}\exp(-(z^c_{ij}-u_{1,l(1,i,j)}-\cdots-u_{P,l(P,i,j)})^2/2\sigma^2_\epsilon)\right)} \times\frac{\left[\Phi(UL^c_{yj}/v_j)-\Phi(LL^c_{yj}/v_j)\right]}{\left[\Phi(UL^*_{yj}/v_j)-\Phi(LL^*_{yj}/v_j)\right]} \tag{3.19}$$

A new value of δ is proposed for each unique y_j of each rank assessment j.

Each κ_j can be updated using a Metropolis-Hastings step because it is not convenient to sample directly from its full conditional distribution. As recommended by Johnson et al. (2002), a lognormal distribution depending on a tuning parameter c_j^2 is used for the proposal distribution of κ^*_j given the current state κ^c_j.

$$q(\kappa^*_j|\kappa^c_j) = \frac{\exp(-(\log(\kappa^*_j)-\log(\kappa^c_j))^2/(2c_j^2))}{\kappa^*_j\sqrt{2\pi c_j^2}}1(\kappa^*_j>0)$$

The lognormal distribution is appealing as a proposal distribution because it is easy to sample from and obeys the restriction that κ_j be positive. The acceptance probability for setting $\kappa_j^{(m)} = \kappa^*_j$ is

$$\min\left(1,\quad\left[\prod_{i=1}^{C(j)-1} p_{(ij)}(\kappa^*_j)/p_{(ij)}(\kappa^c_j)\right](\kappa^*_j/\kappa_j^{(m-1)})^{a_j}\exp(-b_j(\kappa^*_j-\kappa^c_j))\right). \tag{3.20}$$

Recall that $\boldsymbol{\sigma} \equiv (\sigma^2_\epsilon, \sigma^2_{u,1}, \ldots, \sigma^2_{u,P})$. The prior for $\boldsymbol{\sigma}$ is a Dirichlet density, but the full conditional is much more complicated. A Metropolis-Hastings step is used to update these variance parameters. The Dirichlet family of distributions can be used for the proposal distribution. One advantage is that each proposal will satisfy the modeling constraint that the sum of these variance parameters must equal 1. The

transition kernel depends on two tuning parameters, $a_{MH} > 0$ and $b_{MH} \geq 0$.

$$q(\boldsymbol{\sigma}^*|\boldsymbol{\sigma}^c) = Dirichlet[\boldsymbol{\sigma}^*; a_{MH}\boldsymbol{\sigma}^c + b_{MH}\mathbf{1}]$$

The shorthand notation $Dirichlet[\mathbf{x}; \boldsymbol{\alpha}]$ is used to represent the pdf of the Dirichlet distribution as a function of $\mathbf{x}$ and having parameter vector $\boldsymbol{\alpha}$. The general idea is for the proposal distribution to have a mean that is close to the current value. The larger the value of a_{MH} is, the tighter the proposal distribution is. Positive values of b_{MH} essentially shrink each proposed value towards $1/(P+1)$, with the shrinkage more pronounced as b_{MH} increases.

The acceptance probability for $\boldsymbol{\sigma}^*$ is quite involved, so the convention previously undertaken of explicitly stating the acceptance probability in a specific form is now interrupted. The acceptance probability is the minimum of one and the quantity given by Equation 3.21.

$$\frac{q(\boldsymbol{\sigma}^c|\boldsymbol{\sigma}^*)\pi(\boldsymbol{\sigma}^*)\pi(\mathbf{z}|\mathbf{u}_1,\ldots,\mathbf{u}_p,\boldsymbol{\sigma}^*)\prod_{p=1}^{P}\pi(\mathbf{u}_p|\boldsymbol{\sigma}^*)}{q(\boldsymbol{\sigma}^*|\boldsymbol{\sigma}^c)\pi(\boldsymbol{\sigma}^c)\pi(\mathbf{z}|\mathbf{u}_1,\ldots,\mathbf{u}_p,\boldsymbol{\sigma}^c)\prod_{p=1}^{P}\pi(\mathbf{u}_p|\boldsymbol{\sigma}^c)}. \tag{3.21}$$

The random effects can be individually updated using their complete conditional distributions. For each level of each random effect, $u^{(m)}_{p,l(p,i,j)}$ is sampled from the normal distribution with mean μ and variance σ^2, where

$$\sigma^2 = \left[1/\sigma^2_{u,p} + \sum_{i',j',t':l(p,i',j')=l(p,i,j)} 1/\sigma^2_\epsilon\right]^{-1} \tag{3.22}$$

and

$$\mu = \sum_{i',j',t':l(p,i',j')=l(p,i,j)} (\sigma^2/\sigma^2_\epsilon)\left(z_{i'j't'} - \sum_{p'\neq p} u_{p',l(p',i',j')}\right). \tag{3.23}$$

An alternative is to update all random effects in a block using the complete conditional distribution, a multivariate normal. This approach might be preferrable if the complete conditional of the random effects block indicates substantial corre-

lation among the random effects. On the other hand, a block update might not be advisable if the dimensionality of the random effects block is fairly large because of computational difficulties with using large covariance matrices.

Although the structure for the MCMC algorithm has been discussed, there are several implementation considerations that merit attention. Gelman et al. (2004, ch. 10–11) discussed posterior sampling issues and provided many useful hints, which we recommend to the interested reader. We now discuss some of these considerations as they apply to the proposed model. They include selecting the length of the burn-in period, the total number of iterations, and the tuning parameters for the Markov chain.

The burn-in period is an initial fraction of the overall MCMC run that is not used for inference because draws from the Markov chain might not yet be reasonable representations of draws from the actual posterior distribution (see Gelman et al. 2004, p. 295). The burn-in period be sufficiently large. One simple way to make this judgment involves examining trace plots of model parameters and latent variables. The trace plots should reflect stationary behavior beyond the burn-in period if the chain is mixing well.

The total number of iterations should be sufficiently large that posterior inference is not overly dependent on the actual run used (see Gelman et al. 2004, p. 277). Provided the chain mixes well and all parameters are identifiable, a large number of iterations should yield relatively little variability due to the Monte Carlo approximation of the true posterior. If the chain does not mix well, even more iterations will be necessary. Subsampling methods on an initial chain or parallel chains can be used to inform decisions on the number of draws necessary to achieve a desired reduction in Monte Carlo variability. Informally, autocorrelation function plots can be used to see at what lag parameters and latent variables no longer have any correlation worth

noting. This lag might serve as an indication of the minimum number of iterations needed between two draws for quasi-independence of the draws to be a reasonable assumption.

Most of this model's Metropolis or Metropolis-Hastings update steps involve one or more tuning parameters, with $\mathbf{z}$ being the lone exception. Unlike model parameters, the tuning parameters are of no interest other than on the algorithm's behavior. The parameters affect acceptance rates for the proposal distributions and thus are influential. If the tuning parameters are not suitable, the acceptance rates might be too high or too low, both of which can have detrimental effects on chain mixing. We suggest choosing tuning parameters that will yield acceptance rates of 35–45% for each τ_j and κ_j. We also suggest choosing tuning parameters that will yield acceptance rates of 20–30% for $\boldsymbol{\sigma}$. This is not unlike the recommendation by Gelman et al. (2004, p. 307) to target 20% acceptance for vector updates and 40% for scalar updates. It is difficult to specify at the outset which tuning parameter values will give the desired acceptance rates. Running the algorithm many times, each time adjusting the tuning parameters, can help in identifying appropriate values for the tuning parameters. Alternatively, the algorithm may be run one time with tuning parameters initially updated through some part of the burn-in period and then left constant (p. 307).

Unfortunately, the joint binomial and rank response model has several characteristics that predispose the chain to poor mixing. For example, Johnson et al. (2002) used a coupling scheme to determine that in their application of a rank response model posterior draws were quasi-independent if they were 40,000 draws apart, though the authors noted that this number was likely larger than necessary (p. 12). The latent variables for rank data tend to mix poorly if many subjects are ranked because they are subject to order constraints and are not updated as a single block. Despite the use of a modified version of Cowles' (1996) algorithm, the cutpoint parameters still

appear to mix very slowly if there are many random effects. Because of this, the burn-in period and total number of iterations should both be very large. Given the large number of iterations required, thinning of the posterior draws might be necessary if the latent variables are to be saved.

C. Model Extension Allowing Assessment-specific Error Variances

Although the model from Section A is flexible in that it can accommodate an arbitrary number of random effects, it makes a strong assumption in assuming that the variance of the error terms, σ_ϵ^2, is the same across assessments. Because each latent variable has a marginal variance of 1, σ_ϵ^2 represents the proportion of the total marginal variance attributable to error. The constant error variance implies that the relative contribution of the random effects to the total (marginal) variability is assumed to be identical across assessments. This assumption would be in doubt if some assessments are believed to be more strongly related to the random effects than others are.

An extension of the joint binomial and rank response model presented earlier assumes each assessment j has its own error variance σ_j^2. The extension can be handled quite seamlessly. The specifications of the extended model are now briefly noted. Because of its large overlap with the original model, fewer explanory statements are included. The likelihood is completely unchanged from Equation 3.2. The prior distribution is assumed to factor as

$$\begin{aligned}\pi(\mathbf{z}, \mathbf{u}_1, \ldots, \mathbf{u}_P, \boldsymbol{\sigma}, \sigma_1^2, \ldots, \sigma_J^2, \boldsymbol{\kappa}, \boldsymbol{\tau}) = \pi(\mathbf{z}|\mathbf{u}_1, \ldots, \mathbf{u}_P, \boldsymbol{\sigma}, \sigma_1^2, \ldots, \sigma_J^2)\pi(\mathbf{u}_1, \ldots, \mathbf{u}_P|\boldsymbol{\sigma}) \\ \times \quad \pi(\sigma_1^2, \ldots, \sigma_J^2|\boldsymbol{\sigma})\pi(\boldsymbol{\sigma})\pi(\boldsymbol{\tau})\pi(\boldsymbol{\kappa}).\end{aligned} \tag{3.24}$$

A Dirichlet prior is still assumed on $\boldsymbol{\sigma}$, which has the same elements as in the previous model: $(\sigma_\epsilon^2, \sigma_{u,1}^2, \ldots, \sigma_{u,P}^2)$. The assumed prior for $\mathbf{z}|\{\mathbf{u}_1, \ldots, \mathbf{u}_P, \boldsymbol{\sigma}, \sigma_1^2, \ldots, \sigma_J^2\}$ is

nearly identical to Equation 3.6, with the only difference being that σ_j^2 is substituted for every instance of σ_ϵ^2. The prior distributions for $\{\mathbf{u}_1, \ldots, \mathbf{u}_P\}|\boldsymbol{\sigma}$ (Equation 3.7) and $\boldsymbol{\sigma}$ (Equation 3.8) are unchanged.

Before discussing the priors for $\boldsymbol{\kappa}$ and $\boldsymbol{\tau}$ in the extended model, it is important to consider the σ_j^2's. The prior should be chosen so that $E(\sigma_j^2|\boldsymbol{\sigma}) = \sigma_\epsilon^2$ because this condition may be used to help constrain the marginal variance of each z_{ijt} to equal 1, exactly as it does in the original model, provided that the marginal variance is finite. This is easily demonstrated using two well known identities for the marginal mean and marginal variance in terms of conditional means and conditional variances (see, e.g., Gelman et al. 2004, p. 23–24). First note that $z_{ijt}|\{\boldsymbol{\sigma}, \sigma_j^2\} \sim N(0, \sum_p \sigma_{u,p}^2 + \sigma_j^2)$. Equations 3.25 and 3.26 hold if each component quantity exists and is finite.

$$E(z_{ijt}) = E(E(z_{ijt}|\boldsymbol{\sigma}, \sigma_j^2)) = E(0) = 0 \tag{3.25}$$

$$\begin{aligned} Var(z_{ijt}) &= E(Var(z_{ijt}|\boldsymbol{\sigma}, \sigma_j^2)) + Var(E(z_{ijt}|\boldsymbol{\sigma}, \sigma_j^2)) \\ &= E(\sum_p \sigma_{u,p}^2 + \sigma_j^2) + Var(0) = E(\sum_p \sigma_{u,p}^2 + \sigma_\epsilon^2 + (\sigma_j^2 - \sigma_\epsilon^2)) = 1 \end{aligned} \tag{3.26}$$

The last equality in Equation 3.26 follows because the Dirichlet prior on $\boldsymbol{\sigma}$ implies $E(\sum_p \sigma_{u,p}^2 + \sigma_\epsilon^2) = 1$ and because the assumed prior on $\sigma_j^2|\boldsymbol{\sigma}$ implies $E(\sigma_j^2 - \sigma_\epsilon^2) = 0$. It is worth emphasizing that for both the original and extended models, $E(z_{ijt}) = 0$ and $Var(z_{ijt}) = 1$, thus facilitating comparisons between models. However, in the original model the marginal prior distribution of z_{ijt} is normal, while it is nonnormal in the extended model.

We model the assessment-specific error variances using the inverse gamma distribution for convenience in posterior inference and because it is easy to specify a sufficient condition for the marginal variance of each z_{ijt} to equal 1. The $\sigma_j^2|\boldsymbol{\sigma}$ are assumed to be independent and identically distributed with an $IG(b+1, b\sigma_\epsilon^2)$ distri-

bution.

$$\pi(\sigma_1^2, \sigma_2^2, \ldots, \sigma_J^2|\boldsymbol{\sigma}) = \prod_{j=1}^{J} \left((b\sigma_\epsilon^2)^{b+1}(\sigma_j^2)^{-(b+2)} \exp(-b\sigma_\epsilon^2/\sigma_j^2)/\Gamma(b+1)\right) \tag{3.27}$$

For any choice of $b > 0$, the conditional mean of $\sigma_j^2|\boldsymbol{\sigma}$ will equal σ_ϵ^2 and, of ultimate importance, the marginal variance of z_{ijt} (using the entire model framework) will equal 1. For example, in the case where no random effects are present in the model and thus $\sigma_\epsilon^2 \equiv 1$, the marginal prior distribution of z_{ijt} is $t_{2b+2}(0, s^2 = b/(b+1))$ when using Equation 3.27 to specify the prior distribution of the assessment-specific conditional error variances. While the marginal mean and marginal variance will equal 0 and 1, respectively, for any $b > 0$, we recommend that the prior not be too dispersed to limit the nonnormality of the marginal distribution of the latent variables and to make posterior inferences across different assessments more comparable. A recommendation for selecting the value of b is to choose it so that $Pr(2/3 < (\sigma_j^2/\sigma_\epsilon^2) < 3/2)$ equals a desired (large) probability, such as 0.8 (b=9.782) or 0.9 (b=16.082).

The assumed prior distributions for $\boldsymbol{\kappa}$ and $\boldsymbol{\tau}$ have the same form as previously (Equations 3.9 and 3.10). Though not recommended due to the increased model complexity, these priors could be changed because the marginal distribution of each z_{ijt} is now nonnormal. The original priors may be justified in the extended model by noting that a peaked prior on the assessment-specific variances limits the degree of nonnormality, and the marginal location and scale of each z_{ijt} are unchanged.

The MCMC algorithm for posterior sampling is only mildly affected by the suggested model extension to allow assessment-specific error variances. Updating $(\boldsymbol{\tau}, \mathbf{z})$ uses the same approach as for the original model, but replaces σ_ϵ^2 with σ_j^2 in each of Equations 3.13, 3.14, and 3.19, and in the proposal distribution described immediately before Equation 3.15. Updating $\boldsymbol{\kappa}$ requires no changes to the approach described for

the original model. Updating $\boldsymbol{\sigma}$ in the extended model has the same procedure as in the original model, but the acceptance probability is now the minimum of one and the quantity given by Equation 3.28.

$$\frac{\pi(\boldsymbol{\sigma^*})\pi(\sigma_1^2,\ldots,\sigma_J^2|\boldsymbol{\sigma}^*)\prod_{p=1}^{P}\pi(\mathbf{u}_p|\boldsymbol{\sigma}^*)q(\boldsymbol{\sigma}^c|\boldsymbol{\sigma}^*)}{\pi(\boldsymbol{\sigma}^c)\pi(\sigma_1^2,\ldots,\sigma_J^2|\boldsymbol{\sigma^c})\prod_{p=1}^{P}\pi(\mathbf{u}_p|\boldsymbol{\sigma}^c)q(\boldsymbol{\sigma}^*|\boldsymbol{\sigma}^c)} \tag{3.28}$$

An additional step is then added to update each σ_j^2 using its full conditional, which is an $IG(a_j^c, b_j^c)$.

$$a_j^c = b + 1 + \sum_{i=1}^{I}\sum_{t=1}^{T_j} 1/2$$

$$b_j^c = b\sigma_\epsilon^2 + \sum_{i=1}^{I}\sum_{t=1}^{T_j}(z_{ijt} - u_{1,l(1,i,j)} - \cdots - u_{P,l(P,i,j)})^2/2$$

Finally, the random effects are updated using a normal distribution with variance and mean as in Equations 3.22 and 3.23 but with each occurrence of σ_ϵ^2 replaced by σ_j^2.

D. Model Comparison

To this point, it has been assumed that both the number of random effects (P) and the levels of each ($l(p,i,j)$) are known. Nonetheless, primary interest often centers on identifying which random effects are needed in the model. In addition, one may wish to determine if the assumption of a common error variance for all assessments is reasonable or if it should be relaxed. Ideally Bayes factors would be used to compare all models. The difficulty in computing Bayes factors may prohibit their usage, especially because of the large number of latent variables in the model. However, the models have been constructed so that pivotal quantities may be used to compare them. The pivotal quantities considered are not explicit functions of the data, but because they are functions of unknown quantities with a completely specified nominal distribution we refer to them as pivotal quantities.

As in Yuan and Johnson (2011), we consider discrepancy measures formed by binning appropriately standardized residuals into one of B partitions of the real number line. Let each standardized residual r_{ijt} be defined as

$$r_{ijt} \equiv \frac{z_{ijt} - u_{1,l(1,i,j)} - u_{2,l(2,i,j)} - \cdots - u_{P,l(P,i,j)}}{\sigma_j} \tag{3.29}$$

with restrictions made on terms as necessary. For example, $\sigma_j \equiv \sigma_\epsilon$ if a model does not allow assessment-specific error variances, and $u_{p,l(p,i,j)} \equiv 0$ if a model does not allow for the p^{th} random effect. Note that the standardized residuals are computed for each MCMC iteration after the burn-in period using the iteration-specific values of each unknown. For each model considered, the standardized residuals from any given MCMC iteration after convergence are independent and identically distributed as standard normals assuming the model is correct. Thus, these standardized residuals can be conveniently used in forming a discrepancy measure to assess model adequacy. Like Yuan and Johnson (2011), we bin standardized residuals to get observed and expected counts and then form a discrepancy measure based on Pearson's goodness-of-fit statistic. We suggest choosing as bin thresholds $(\Phi^{-1}((b-1)/B), \Phi^{-1}(b/B)]$, $b = 1, \ldots, B$ so that each bin is equally likely if the model is correct. In order to find a model weakness, observed bin counts are compared with expected bin counts, but the comparisons must be done so as to give the discrepany measure a good chance to be large if an important random effect is not included or if the conditional error variances are incorrectly assumed to be constant across assessments.

First consider that the p^{th} random effect is important but has been omitted from the model (or equivalently, $u_{p,l(p,i,j)}$ has been constrained to equal zero for each level l). Following the logic explained by Yuan and Johnson (2011), the absence of this effect should make the standardized residuals for the l^{th} level of this effect tend to be more (less) than zero as the true effect $u_{p,l(p,i,j)}$ is positive (negative). Recall that

L_p represents the number of levels in the p^{th} random effect, and define $O_{p,l,b}$ as the observed number of standardized residuals for the l^{th} level of the p^{th} random effect that are in bin b. Likewise, define $E_{p,l,b}$ as the expected number in the bin. The discrepancy measure D_p systematically compares the observed and expected counts at each level of the random effect and is properly viewed as an adaptation of the discrepancy measure proposed by (Yuan and Johnson 2011, p. 9).

$$D_p \equiv \sum_{l=1}^{L_p} \sum_{b=1}^{B} \frac{(O_{p,l,b} - E_{p,l,b})^2}{E_{p,l,b}} \tag{3.30}$$

This type of discrepancy measure is a pivotal quantity that has an approximate $\chi^2(L_p(B-1))$ nominal distribution when all expected counts are sufficiently large (p. 9). We emphasize that the discrepancy measure is computed separately for each post-burn-in MCMC iteration.

Similarly, to check the assumption that $\sigma_j^2 \equiv \sigma_\epsilon^2$ for each assessment, let $O_{j,b}$ ($E_{j,b}$) be the observed (expected) number of residuals from the j^{th} assessment that are in the b^{th} bin. The discrepancy measure D_{VAR}, defined in Equation 3.31, has an approximate $\chi^2(J(B-1))$ distribution if the model is correct and each $E_{j,b}$ is sufficiently large.

$$D_{VAR} \equiv \sum_{j=1}^{J} \sum_{b=1}^{B} \frac{(O_{j,b} - E_{j,b})^2}{E_{j,b}} \tag{3.31}$$

An alternative is to group the squared standardized residuals into bins with thresholds taken from quantiles of the $\chi^2(1)$ distribution, which is similar to D_{VAR} but ignores the sign of the standardized residual.

Choice of the number of bins should be governed by the ability to have an adequate expected count in each bin. For random effects with many levels, this might mean it is necessary to use fewer bins. However, if the number of bins is too small, the discrepancy measure might not have adequate ability to detect a model

violation. We recommend four or five bins, depending on the number of residuals that will be binned. In an unbalanced design, it is possible that some of the inner sums in D_P or D_{VAR} involve expected counts that are too small for the selected number of bins. One consideration would be to have the number of bins vary across levels/assessments, but if the number of bins would be too small it might be advisable to simply ignore the sum for levels/assessments where the chi-square approximation is in doubt. Of course, either the decision to let the number of bins vary or to exclude particular summands would affect the degrees of freedom for the discrepancy measure's approximate distribution.

Another important choice is which quantiles to use from the reference distribution and the collection of post-burn-in discrepancy measures; the reference quantile and sample quantile are both used in bounding the PPP p-value. For the reference quantile, we recommend using (only) the 99^{th} percentile for the reference distribution because if a lower percentile is used then the upper bound on the PPP p-value cannot be less than 0.01. This is not absolutely necessary though; because it is an upper bound on the p-value and not the p-value itself it is reasonable to consider other values than, say, 0.01 or 0.05 as indicating model inadequacy (see (Yuan and Johnson 2011, pp. 11-12)). Johnson (2007) noted that care must be taken when using very large quantiles if the reference distribution is only an approximation to the true distribution. For this reason, we recommend that the theoretical quantile not exceed the 99^{th} percentile. For a given reference quantile, the procedure of Yuan and Johnson (2011) would select the sample quantile from the collection of discrepancy measures by choosing the smallest one that still exceeds the reference quantile. We recommend this same approach to selecting the sample quantile.

CHAPTER IV

APPLICATION

This chapter applies the previously described model to data on primate intelligence. Section A contains information on the data. Among many candidate models for these data, one is selected based on fitting progressively more complex models until the model is deemed adequate. Such a determination relies on comparing pivotal quantities designed to pick up deviations from the model for the latent variables in $\mathbf{z}$. The model selection is described in Section B. Detailed results for the selected model are presented in Section C. Section D investigates sensitivity of the selected model to modest changes in the prior distribution and to the exclusion of some data. Section E discusses the conclusions in the context of previous research on primate intelligence.

A. Primate Intelligence Data

Primate intelligence has been studied with the goals of identifying which species are most intelligent and plausibly explaining characteristics of species that might explain their intelligence. Because intelligence is a latent trait that is not directly measured, various tests have been constructed to assess intelligence. In light of this, Spearman (1904) investigated "general intelligence" in humans; his analysis is widely credited with introducing the use of factor analysis as a method to understand one or more latent traits based on a multivariate response vector.

Many different tests have been used for studying primate intelligence. These tests have been referred to as *procedures*, and a group of related procedures comprise what is called a *paradigm* (Johnson et al. 2002). Because of the diversity in procedures and paradigms, it is plausible to suppose that more than one latent trait might be appropriate to explain performance.

A collaborator, Federica Amici, has personally collected primate data on many assessments; other data were generously shared with us (see the Acknowledgments immediately preceding the Table of Contents). We use the term *assessments* rather than *procedures* or even *tasks* because the assessments differ in level of uniqueness; some are quite unrelated to all others while some are only different conditions of a common task. The data comprise performance results of 100 primates, distributed unevenly across seven species: chimpanzees, bonobos, gorillas, orangutans, spider monkeys, capuchin monkeys, and long-tailed macaques. Twenty assessments were used representing six (narrowly defined) paradigms. Some assessments were not performed among all species. Table II contains the classifications used for each assessment's response type and paradigm. Table III contains counts of the number of animals assessed from each species.

The first paradigm concerns inhibition and encompasses the first five assessments of Table II. For the first two assessments, performance was scored using the ratio of the percentage of correct trials in the experimental condition to the percentage of correct trials in the control condition. The use of the ratio does not lend itself to a binomial response, so the rank of the ratio is used as the response. The third assessement was the number of correct trials in two trials and was modeled as a binomial response. The fourth assessment was a measurement in seconds of the amount of time the animal was able to delay gratification, and because of potential skewness in the data it was sensible to treat the data in terms of their ranks. The fifth assessment was the number of correct trials out of ten and was modeled as a binomial response. The data from these five assessments for monkeys (spider monkeys, capuchin monkeys, and long-tailed macaques) and great apes (chimpanzees, bonobos, gorillas, and orangutans) were previously analyzed by Amici et al. (2008). The monkey data for these five assessments were collected by Amici et al. (2008), as were the great ape data

Table II. The 20 assessments used for studying primate intelligence. Further information on most assessments can be found in earlier published work, as described in the text.

No.	Assessment Description	Modeled	Paradigm
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
18	Gaze Following Ceiling	Rank	Gaze Following
19	Gaze Following Barrier	Rank	Gaze Following
20	Reversed Contingency	Rank	Rev. Contingency

Table III. The number of animals from each species observed for each assessment. One hundred unique animals were observed in the combined data set.

	Number of Animals Assessed						
Assessment (Paradigm)	Chim-panzee	Bonobo	Gorilla	Orang-utan	Spider Monkey	Capuchin Monkey	Long-tailed Macaque
1 (1)	7	4	7	6	15	19	12
2 (1)	7	4	7	6	17	19	12
3 (1)	8	4	6	6	14	16	12
4 (1)	5	5	4	8	12	12	12
5 (1)	6	4	6	7	18	28	12
6 (2)	11	5	8	10	14	12	12
7 (2)	7	5	4	8	14	12	12
8 (3)	7	4	7	6	13	12	12
9 (3)	7	4	7	6	13	12	12
10 (3)	7	4	7	6	13	12	12
11 (3)	7	4	7	6	13	12	12
12 (4)	18	5	5	5	12	12	12
13 (4)	18	5	5	5	12	12	12
14 (4)	18	5	5	5	12	12	12
15 (4)	17	5	5	5	12	12	12
16 (4)	17	5	5	5	12	12	12
17 (4)	17	5	5	5	12	12	12
18 (5)	0	0	0	0	13	12	0
19 (5)	0	0	0	0	13	12	0
20 (6)	0	0	0	0	12	12	12
Overall	19	5	8	10	18	28	12

from the first three assessments and the gorilla and orangutan data from the fourth assessment. The bonobo and chimpanzee data for the fourth assessment were from Rosati et al. (2007). All great ape data for the fifth assessment were from Vlamings et al. (2010).

The second paradigm involves memory and used two assessments, one involving three trials with a 30 second period and the other involving three trials with a 30 minute period. Both of these were modeled as binomial responses. Data on monkeys were collected and analyzed by Amici et al. (2010), while data on great apes were collected and analyzed by Barth and Call (2006) and Amici et al. (2010).

The third paradigm tests the ability to deal with transposition. Four assessments were used, each corresponding to a different condition for the same task. Each assessment consisted of either one or two trials and was modeled as a binomial response. As with data used in the second paradigm, data on monkeys were originally collected and analyzed by Amici et al. (2010); data on great apes were collected and analyzed by Barth and Call (2006) and also analyzed by Amici et al. (2010).

The fourth paradigm consists of assessments designed to test the ability to perceive connections. Six assessments were used, each corresponding to a different condition of a common task. Each assessment used six trials and was modeled as a binomial response. Data on monkeys were collected by Amici, with the cloth conditions having been analyzed by Amici et al. (2010). Data on great apes were collected and analyzed by Herrmann et al. (2008) and also analyzed by Amici et al. (2010).

The fifth paradigm is represented by two assessments dealing with gaze following, each assessment using a different experimental and control condition of a particular task. The outcome for these two assessments used the ratio of looks at a particular area in the assessment's experimental condition to the combined number of looks in the assessment's experimental and control conditions. The ratios were modeled

in terms of their ranks. Data are only available on spider monkeys and capuchin monkeys; the data were collected and analyzed by Amici et al. (2009).

The sixth paradigm is represented by a single assessment because of its uniqueness compared to the other nineteen assessments. The original response for this assessment consisted of the percent of correct responses in the experimental condition. However, the number of trials differed across animals and was dependent on performance. The percentages were converted to ranks for modeling purposes. We prefer the rank modeling for this assessment in part because the number of trials was related to performance, but also because the number of trials might have been much larger than for any other assessment making this assessment very influential. Amici collected these data on the three aforementioned monkey species, but these data have not yet been used in published results.

All of the memory, transposition, and support assessments could be considered to come from a single paradigm, but we split them up because there are a sufficient number of assessments in each component to model them separately and because if combined there are very few paradigms with which to assess paradigm-specific effects. The results suggest that this split is justified because there are difference between relative species performance that are not consistent across these three narrowly defined paradigms.

B. Model Comparison

The primate intelligence data served as a motivating application for developing the joint model while allowing for random effects. The data set is especially informative because it permits investigation of several important questions.

- Is there substantial interanimal variability in intelligence after accounting for

species effects? If so, how does interanimal variability compare to interspecies variability (on a latent scale)?

- Do assessments from different paradigms yield different conclusions in comparing intelligence between or within species?
- Does the relative contribution of error (i.e., the component of latent performance variables not explained by species- or animal-related effects) differ across assessments?

The data have both rank and binomial assessments, so the joint model from Chapter III is used. Many models are considered, but before discussing the differences between various models we lay the framework common to these models using the notation from Chapter III. Several subscripts are used: $i = 1, \ldots, 100$ represents the subject, or animal, number; $j = 1, \ldots, 20$ represents the assessment number; and $t = 1, \ldots, T_j$ represents the trial number for (binomial) assessment j. The observed response y_{ijt} is binary for binomial assessments with $y_{ijt} = 1$ denoting a success for the trial. The observed response for rank assessments is in the interval $[1, C(j)]$, with $y_{ijt} = 1$ representing the worst performance among the $C(j)$ subjects ranked for assessment j and with $y_{ijt} = C(j)$ the best (if unique). This ranking is in the opposite direction of what might have been expected, but it was chosen so that higher values of y_{ijt} are better regardless of whether assessment j has binomial or rank responses. The value of $B(j)$ is binary and equal to one if assessment j uses a binomial response, while it is equal to zero if assessment j uses a rank response.

The likelihood for the data depends on a vector of latent variables $\mathbf{z}$, cutpoint parameters $\boldsymbol{\tau}$, and parameters influencing the probability of two rankings being tied $\boldsymbol{\kappa}$. For convenience, part of the likelihood is defined using the function $p_{(ij)}(\kappa_j)$ stated in Equation 4.1. This function is used only for rank data. The notation $z_{(i),j}$ $(y_{(i),j})$

represents the i^{th} smallest latent variable (observed rank) for the $C(j)$ animals ranked in assessment j. Any animal not ranked for assessment j is ignored.

$$p_{(ij)}(\kappa_j) = \begin{cases} \exp(-(z_{(i+1),j} - z_{(i),j})/\kappa_j) & \text{if } y_{(i+1),j} = y_{(i),j} \\ 1 - \exp(-(z_{(i+1),j} - z_{(i),j})/\kappa_j) & \text{if } y_{(i+1),j} > y_{(i),j}. \end{cases} \tag{4.1}$$

It is possible to model each y_{ijt}, $i = 1, \ldots, 100$, $j = 1, \ldots, 20$, $t = 1, \ldots, T_j$ by assuming that there is no systematic pattern to the missingness of $\mathbf{y}$. Chapter III illustrates how this is done. Such an implementation would be inefficient because there are many missing values of y_{ijt}, and the missing values do not affect the likelihood. We modeled only observed y_{ijt} values. However, we present the likelihood (Equation 4.2), prior, and MCMC algorithm in the more general form to avoid complexities in product and summation index ranges for missing data. The variable w_{ijt} is an indicator of whether y_{ijt} was observed ($w_{ijt} = 1$) or not ($w_{ijt} = 0$).

$$\begin{aligned} f(\mathbf{y}|\mathbf{z}, \boldsymbol{\tau}, \boldsymbol{\kappa}) = & \left[\prod_{j:B(j)=1} \prod_{i=1}^{100} \prod_{t=1}^{T_j} (1(\{y_{ijt} = 0 \cap z_{ijt} \leq \tau_j\} \cup \{y_{ijt} = 1 \cap z_{ijt} > \tau_j\}))^{w_{ijt}} \right] \\ & \times \left[\prod_{j:B(j)=0} \left(\prod_{i=1}^{C(j)-1} p_{(ij)}(\kappa_j) \right) \left(\prod_{i=1}^{100} \prod_{i':z_{i'j}<z_{ij}} (1(y_{i'j} \leq y_{ij}))^{w_{ij}w_{i'j}} \right) \right] \end{aligned} \tag{4.2}$$

The form of the likelihood is unchanged for all models considered, so the real differences between models are introduced by assuming different prior distributions for the latent variables $\mathbf{z}$. Up to four random effects are considered: species effects, animal effects, species*paradigm effects, and animal*paradigm effects. Eschewing the generic but easily generalized notation of Chapter III 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 animal characteristic, $s(i) = 1, \ldots, 7$ can be used to represent the species of animal i and $\theta_{s(i)}$ can be used to represent the species effect. The values of $s(i)$ from $s(i) = 1$ to $s(i) = 7$ correspond to chimpanzees, bonobos, gorillas, orangutans, spider monkeys, capuchin monkeys, and long-tailed macaques, respectively. The animal effect for animal i is represented by γ_i. Each assessment belongs to one of six paradigms, so $g(j) = 1, \ldots, 6$ can be used to represent the paradigm for assessment j. The values of $g(j)$ from $g(j) = 1$ to $g(j) = 6$ correspond to inhibition, memory, transposition, support, gaze following, and reversed contingency, respectively. The species*paradigm effect is represented by $\omega_{s(i),g(j)}$. The animal*paradigm effect is represented by $\eta_{i,g(j)}$.

The random effects combine with the assessment error variance to parameterize the prior distribution for the latent variables. The most complex model that we considered assumes that the latent variables comprising $\mathbf{z}$ have assessment-specific error variances $\sigma_j^2 : j = 1, \ldots, 20$ and that conditional on the error variances and random effects the z_{ijt} are mutually independent with a $N(\theta_{s(i)} + \gamma_i + \omega_{s(i),g(j)} + \eta_{i,g(j)}, \sigma_j^2)$ distribution. Modifications of this model consist of removing random effects and forcing the error variances to be common across models (i.e., σ_ϵ^2 in place of σ_j^2). The models considered appear in Table IV.

Although many of the models appear to be nested within other models, this is often not the case because non-local priors are used on the variance parameters of the random effects. Nonetheless, it is clear that each model can be considered as an extension of the simplest model $M_{0,0}$. Likewise, each model can be considered a simplification of the most complex model $M_{SPAP,J}$. Before stating the prior specification for these models, the rationale for considering them is presented. Model $M_{0,0}$ is the simplest model that could be considered as it has no random effects (and consequently the conditional variance is set to equal one). The most natural extension to

Table IV. Models under consideration for the conditional mean and variance of the latent variable z_{ijt}. In each model the z_{ijt}'s are assumed to be conditionally independent and have conditionally normal distributions.

	Latent Variables	
Model	Conditional Mean	Conditional Variance
$M_{0,0}$	0	1
$M_{S,0}$	$\theta_{s(i)}$	σ^2_ϵ
$M_{SA,0}$	$\theta_{s(i)} + \gamma_i$	σ^2_ϵ
$M_{SP,0}$	$\theta_{s(i)} + \omega_{s(i),g(j)}$	σ^2_ϵ
$M_{SPAP,0}$	$\theta_{s(i)} + \gamma_i + \omega_{s(i),g(j)} + \eta_{i,g(j)}$	σ^2_ϵ
$M_{S,J}$	$\theta_{s(i)}$	σ^2_j
$M_{SA,J}$	$\theta_{s(i)} + \gamma_i$	σ^2_j
$M_{SP,J}$	$\theta_{s(i)} + \omega_{s(i),g(j)}$	σ^2_j
$M_{SPAP,J}$	$\theta_{s(i)} + \gamma_i + \omega_{s(i),g(j)} + \eta_{i,g(j)}$	σ^2_j

the model is to allow for species effects. Model $M_{S,0}$ represents a model with species effects and can be compared with $M_{0,0}$ to check for differences in intelligence between species. Model $M_{S,J}$ also allows for species effects, but relaxing the assumption of common conditional error variance allows the species effect to vary in relative importance across assessments. Models $M_{SA,0}$ and $M_{SA,J}$ are particularly interesting because they can be used not only to test for species and animal effects but also provide an indication of the relative importance of each. If animal effects were found to be as or more important than species effects, this could have enormous implications on the interpretation of previous research and the conduct of future research. The combined primate intelligence data set is remarkable because of its large number of observations on the animal level and its inclusion of great apes and monkeys.

Instead of adding an animal effect, another modest extension of the species-effect model might add a species*paradigm interaction effect. Some early model fits suggested this model might be preferable. The 20 assessments used can be grouped into 6 paradigms, and it is possible that the paradigms assess different aspects of intelligence. Johnson et al. (2002) used a similar modeling approach (that served as the basis of our modeling of rank data) with a meta-analysis of primate intelligence rank data and investigated the evidence for genus*paradigm interactions. The comparison of $M_{SP,0}$ with $M_{S,0}$ (or $M_{SP,J}$ with $M_{S,J}$) assesses the need for species*paradigm effects.

Finally, the species effect model can be extended by allowing species*paradigm and animal effects. In this case, it is sensible to also allow for animal*paradigm effects. The models $M_{SPAP,0}$ and $M_{SPAP,J}$ include all of these random effects.

The generic form of the prior for the most complex model, $M_{SPAP,J}$, is stated in Equation 4.3; recall that the general prior structure is discussed in detail in Chap-

ter III.

$$\begin{aligned}
&\pi(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{\eta}, \boldsymbol{\sigma}, \sigma_1^2, \ldots, \sigma_{20}^2, \boldsymbol{\kappa}, \boldsymbol{\tau}) \\
&= \pi(\boldsymbol{\tau})\pi(\boldsymbol{\kappa})\pi(\mathbf{z}|\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{\eta}, \sigma_1^2, \ldots, \sigma_J^2)\pi(\boldsymbol{\sigma}) \\
&\times \pi(\boldsymbol{\theta}|\boldsymbol{\sigma})\pi(\boldsymbol{\gamma}|\boldsymbol{\sigma})\pi(\boldsymbol{\omega}|\boldsymbol{\sigma})\pi(\boldsymbol{\eta}|\boldsymbol{\sigma})\pi(\sigma_1^2, \ldots, \sigma_{20}^2|\boldsymbol{\sigma})
\end{aligned} \tag{4.3}$$

Rather than restate the joint prior distribution for every model considered, we note that this prior specification can be applied to every model if the component prior densities are appropriately chosen.

The vector $\boldsymbol{\sigma}$ contains the variance parameter σ_ϵ^2 followed by the variance of any random effects in the model. For example, with $M_{SPAP,J}$, $\boldsymbol{\sigma}$ is defined as $(\sigma_\epsilon^2, \sigma_\theta^2, \sigma_\gamma^2, \sigma_\omega^2, \sigma_\eta^2)$. 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.

To penalize models for including unnecessary random effects, we restrict attention to values of α that are greater than one. Thus, the prior density is zero whenever any of these variance parameters are zero and the prior is non-local. 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, so the model is penalized for its inclusion. The specific rule used for the primate intelligence analysis is 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 length of $\boldsymbol{\sigma}$ because the expected value of each element is the inverse of

Table V. Values of the hyperparameter α in the prior distribution $\boldsymbol{\sigma} \sim Dirichlet(\alpha \mathbf{1})$ to satisfy the condition that $Pr(\sigma_\theta^2 < E(\sigma_\theta^2)/25) = 0.01$. For the model with only one variance parameter, the prior is degenerate at one.

No. of Elements in $\boldsymbol{\sigma}$	Models	α
1	$M_{0,0}$	1
2	$M_{S,0}$, $M_{S,J}$	1.240
3	$M_{SA,0}$, $M_{SA,J}$, $M_{SP,0}$, $M_{SP,J}$	1.350
5	$M_{SPAP,0}$, $M_{SPAP,J}$	1.425

the number of elements. Because in a given model each variance parameter in $\boldsymbol{\sigma}$ has the same marginal distribution, the criterion could equivalently have been to require that $Pr(\sigma_\epsilon^2 < E(\sigma_\epsilon^2)/25) = 0.01$. The obvious exception to this rule is that for $M_{0,0}$, the only element of $\boldsymbol{\sigma}$ (σ_ϵ^2) has been fixed at one, which means that any value of α may be chosen and that the criterion will never be satisfied, but that is immaterial. Table V denotes the values of α that were used for each model.

Figure 1 shows the marginal density for an arbitrary element of $\boldsymbol{\sigma}$. Note that the marginal density when there is a single variance parameter is not depicted because it is degenerate at 1.

Whenever present in the model, we assume all of the random effects are conditionally independent and have the following prior distributions: $\boldsymbol{\theta}|\boldsymbol{\sigma} \sim N(\mathbf{0}, \sigma_\theta^2 \mathbf{I})$; $\boldsymbol{\gamma}|\boldsymbol{\sigma} \sim N(\mathbf{0}, \sigma_\gamma^2 \mathbf{I})$; $\boldsymbol{\omega}|\boldsymbol{\sigma} \sim N(\mathbf{0}, \sigma_\omega^2 \mathbf{I})$; $\boldsymbol{\eta}|\boldsymbol{\sigma} \sim N(\mathbf{0}, \sigma_\eta^2 \mathbf{I})$. On the other hand, for any of these effects that are not included in the model, the prior distribution for that effect is degenerate at $\mathbf{0}$.

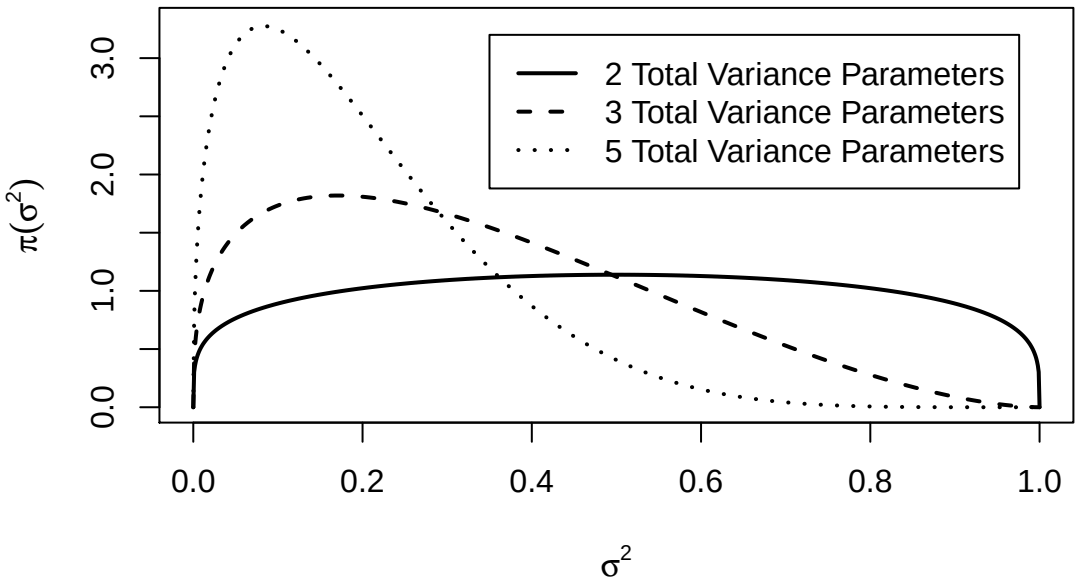

Fig. 1. Marginal distribution of each variance parameter in $\boldsymbol{\sigma}$ as a function of the number of variance parameters. Each variance parameter has a beta marginal distribution. The non-local nature of the prior is apparent from the marginal density going to zero as the variance goes to zero.

If the model has assessment-specific conditional error variances, we assume that conditional on $\boldsymbol{\sigma}$ the σ_j^2's are iid $IG(b+1, b\sigma_\epsilon^2)$. The hyperparameter b is chosen to be 9.782 so that each assessment has an 80% chance under the prior distribution that $\sigma_j^2/\sigma_\epsilon^2 \in (2/3, 3/2)$. While the prior mean of each error variance is the same regardless of $b > 0$, the prior is informative with this value of b because we do not desire dramatic variations among assessments. The scale of the latent variables z_{ijt} is not inherently established by the data, and in our view a very small value of b would only weakly identify the scale of the latent variables, especially if the model's random effects are relatively unimportant. If the model does not have assessment-specific conditional error variances, the prior distribution of each $\sigma_j^2|\boldsymbol{\sigma}$ is degenerate at σ_ϵ^2.

The conditional prior distribution of the latent variable vector $\mathbf{z}$ is given by Equation 4.4.

$$\begin{aligned}&\pi(\mathbf{z}|\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{\eta}, \sigma_1^2, \ldots, \sigma_J^2) \\ &\quad = \prod_{i=1}^{100} \prod_{j=1}^{20} \prod_{t=1}^{T_j} (2\pi\sigma_j^2)^{-1/2} \exp(-(z_{ijt} - \theta_{s(i)} - \gamma_i - \omega_{s(i),g(j)} - \eta_{i,g(j)})^2/2\sigma_j^2)\end{aligned} \tag{4.4}$$

This prior is in the form of $M_{SPAP,J}$ but by properly constraining the σ_j^2's and the random effects it also applies to every model considered.

The final step in setting up the models is to choose the priors for $\boldsymbol{\kappa}$ and $\boldsymbol{\tau}$. The priors selected are the same across models. The κ_j for rank assessments are assumed to be mutually independent and have a $Gamma(1,1)$ (or $Exp(1)$) distribution. The τ_j for binomial assessments are assumed to be mutually independent and have a $Cauchy(0, 0.5^2)$ distribution. Both of these priors were chosen based partially on the actual data set, which is not particularly problematic because (a) these are nuisance parameters, (b) they are common to all models considered, and (c) the priors were still chosen somewhat vaguely in that they are not assessment specific, even though

the posterior distributions would support assessment-specific priors.

In the interest of full disclosure, the process of selecting the values is now stated. Recognizing that every model considered implies $E(z_{ijt}) = 0$ and $Var(z_{ijt}) = 1$, the prior for $\boldsymbol{\tau}$ was selected to be centered at 0 and to be consistent with the range of observed success proportions across the 14 binomial assessments (0.044–0.885). Note that the $(1 - 0.5)/14$ and $(14 - 0.5)/14$ quantiles of the $Cauchy(0, 0.5^2)$ distribution are ± 4.44, which contain virtually all of the probability for a standard normal random deviate. Thus, the prior did not seem overly restrictive because both very small and very large success proportions could reasonably come from such a prior. However, a scale of 0.5 is still an arbitrary choice, even when using observed proportions as a guideline for selecting the scale. In retrospect, we would have preferred a larger value of the scale hyperparameter.

We also informally used observed characteristics of the data to choose the prior for $\boldsymbol{\tau}$. We wanted to find values of κ that would be consistent with the observed proportion of rank values that were unique. There was widespread variation in these proportions. Initially we used prior-predictive simulation under model $M_{0,0}$ to try to approximately reproduce the correct number of ties. The number of animals assessed affects the observed proportion of unique rankings in all models because the order statistics used in determining model tie probabilities tend to become closer to each other when more animals are included, so we randomly drew 20 animals from each assessment, looked at the observed proportions of unique values, and tried to reproduce them. While not precise, we concluded that a value of κ near 1 would suit our purposes, so we chose a $Gamma(2, 2)$ for κ_j's prior distribution. However, early model fits indicated that some κ were very close to zero and some were bigger than 1, so we changed the prior on the κ_j's to be more dispersed but still have mean 1.

Posterior inference used MCMC sampling with the algorithm described in Chapter III. Different alternatives were employed for chain initialization.

Formal assessment of the competing models would ideally use Bayes factors, but computation of the Bayes factors is particularly challenging. The estimator proposed by Gelfand and Dey (1994) (see Equation 2.3) can be difficult to successfully implement because of the difficulty in finding an appropriate importance sampling function that is not dominated by a very small number of values.

Each model in Table IV is fitted to the primate intelligence data set. Model comparison is accomplished by computing discrepancy measures based on pivotal quantities and by comparing posterior inferences from different models. The discrepancy measures as adapted from Yuan and Johnson (2011) were presented generally in Chapter III, and their motivation and characteristics as provided by these authors were also restated there. We turn our attention to specifying the discrepancy measures in a specific form given the candidate models considered. Discrepancy measures are computed for all but the most complex model. For each model, define the standardized residuals r_{ijt} as follows.

$$r_{ijt} \equiv \frac{z_{ijt} - \theta_{s(i)} - \gamma_i - \omega_{s(i),g(j)} - \eta_{i,g(j)}}{\sigma_j} \tag{4.5}$$

Restrictions must be made on terms as necessary. For example, $\sigma_j \equiv \sigma_e$ if the model does not allow task-specific error variances, and $\gamma_i \equiv 0$ if the model does not allow for animal effects. Recall that for each model, the standardized residuals are independent and identically distributed as standard normals assuming the model is correct, and thus functions of the r_{ijt} are pivotal quantities.

In order to have an effective discrepancy measure for use in model comparison, it is important that the discrepancy measure be sensitive to model misspecification. It is therefore useful to select model-specific discrepancy measures that are designed

to pick up features that could be explained by a more general candidate model.

If the error variances are incorrectly assumed to be common across assessments, the standardized residuals for some assessments should have more or less variance than the standardized residuals for other assessments. An appropriate discrepancy measure to detect this misspecification involves binning residuals on an assessment-by-assessment basis. Let $E_{b,j}$ be the expected number of residuals for assessment j in bin b, where $b = 1, 2, \ldots, B$. The bins are defined such that $E_{b,j}$ is the same for each bin b by choosing bin cutpoints of $-\infty = z_0, z_{1/B}, z_{2/B}, \ldots, z_{B/B} = \infty$, where z_α is the lower $(100\alpha)^{th}$ percentile of the standard normal distribution.

$$D_{VAR} = \sum_{j=1}^{J} \left[\sum_{b=1}^{B} \frac{(O_{b,j} - E_{b,j})^2}{E_{b,j}} \right] \tag{4.6}$$

The inner sum in Equation 4.6 has an approximate χ^2_{B-1} distribution for each j provided each expected count $E_{b,j}$ is sufficiently large and the assumed model is correct. Furthermore, if the model is correct the approximate distribution of the overall discrepancy measure is $\chi^2_{J(B-1)}$. Conversely, if the model is incorrect because there are large task-to-task differences in error variability, the inner sums would tend to be larger than $B - 1$ and the overall discrepancy measure would tend to be larger than $J(B - 1)$. We used five bins for this discrepancy measure and excluded any quantities with $E_{b,j}$ less than six because we wanted the chi-square approximation to be reasonable at the 99^{th} percentile of the reference distribution. This meant assessments 18 and 19 were excluded, and the approximate reference distribution is χ^2_{72}.

Because D_{VAR} considers the sign of a standardized residual, it might not be as suited to pick up non-constant error variances. A related discrepancy measure which only considers the magnitude of the standardized residuals, or equivalently considers

the squared value of the standardized residuals, has the same functional form as in Equation 4.6. The only difference is that bin cutpoints are quantiles from the χ^2_1 distribution, the reference distribution for each r^2_{ijt}. The expected and observed bin counts by assessment for the squared residuals would take the place of $E_{b,j}$ and $O_{b,j}$ in Equation 4.6. We refer to this alternative discrepancy measure as $D_{VAR,2}$. Again using five bins that each have the same expected count, and again excluding assessments with expected bin counts of less than six, the approximate reference distribution for $D_{VAR,2}$ is also χ^2_{72}.

The remaining discrepancy measures also excluded any summand based on an expected bin count less than six.

A discrepancy measure for detecting the need for an omitted random effect can be formed by binning residuals according to the levels of the omitted random effect. To assess the need for a species effect to be added to $M_{0,0}$, the discrepancy measure is based on the quantities $E_{b,s}$ and $O_{b,s}$, which are the expected and observed number of standardized residuals in bin b for measurements from species s.

$$D_S = \sum_{s=1}^{S} \left[\sum_{b=1}^{B} \frac{(O_{b,s} - E_{b,s})^2}{E_{b,s}} \right]$$

Five bins were used for this discrepancy measure, and each of the seven species had an expected count in each bin of at least six. The approximate reference distribution is χ^2_{28}.

To assess the need for an animal effect, the discrepancy bins residuals on a per animal basis.

$$D_A = \sum_{i=1}^{I} \left[\sum_{b=1}^{B} \frac{(O_{b,i} - E_{b,i})^2}{E_{b,i}} \right]$$

Five bins were used, but 32 of the 100 animals did not have expected counts of at least six standardized residuals in each bin. The approximate reference distribution

is χ^2_{272}.

In assessing the need for a species*paradigm effect, the discrepancy measure uses expected and observed counts of residuals in bin b, grouped by species s and paradigm g.

$$D_{SP} = \sum_{s=1}^{S} \sum_{g=1}^{G} \left[\sum_{b=1}^{B} \frac{(O_{b,sg} - E_{b,sg})^2}{E_{b,sg}} \right]$$

Four bins were used, and 30 of the 42 species*paradigm combinations had expected counts of at least six standardized residuals. The approximate reference distribution is χ^2_{90}.

A discrepancy measure for testing the need to include animal*paradigm effects could be similarly constructed. However, the data were considered too sparse to reliably test for such an effect. We did not consider any model extensions where the only effect added was an animal*paradigm random effect.

Table VI contains upper bounds on the PPP p-values for testing that the variance for particular random effects is nonzero and that for at least one assessment, $\sigma_j^2 \neq \sigma_\epsilon^2$. The tests were implemented using various discrepancy measures, but in each case the 99^{th} percentile of the reference distribution was used in identifying the upper bound on the p-value. Thus, the minimum upper bound that could be obtained is 0.01. The lower of the PPP p-value bounds resulting from D_{VAR} and $D_{VAR,2}$ is reported for detection of non-constant error variance. Because none of the models suggest a need for the error variances across assessments to differ, the discussion of these models will focus on the models assuming a common error variance.

Each of the models that were considered was fitted using 100,000 burn-in iterations and then 1,500,000 more iterations. Beginning at the simplest model, $M_{0,0}$, it is obvious from Table VI that the model is inadequate because the discrepancy measure designed to pick up species effects tends to be much larger than would be

Table VI. Tests of model inadequacies for the models considered. The bounds on PPP p-values are based on discrepancy measures using pivotal quantities and are upper bounds. The tests are to detect species effects ($\sigma^2_\theta > 0$); animal effects ($\sigma^2_\gamma > 0$); species*paradigm effects (σ^2_ω); and assessment-specific error variances ($\sigma^2_j \neq \sigma^2_\epsilon$). When testing the need for assessment-specific error variances, the lower of the bounds resulting from D_{VAR} and $D_{VAR,2}$ is reported; in every case this was the bound from $D_{VAR,2}$. Each model was fitted using 100,000 burn-in iterations followed by 1,500,000 iterations.

Model	Potential Violation	Bound on p-value	Potential Violation	Bound on p-value	Potential Violation	Bound on p-value
$M_{0,0}$	$\sigma^2_\theta > 0$	0.01				
$M_{S,0}$	$\sigma^2_\gamma > 0$	0.26	$\sigma^2_\omega > 0$	0.01	$\sigma^2_j \neq \sigma^2_\epsilon$	0.73
$M_{S,J}$	$\sigma^2_\gamma > 0$	0.18	$\sigma^2_\omega > 0$	0.01		
$M_{SP,0}$	$\sigma^2_\gamma > 0$	0.82			$\sigma^2_j \neq \sigma^2_\epsilon$	0.61
$M_{SP,J}$	$\sigma^2_\gamma > 0$	0.90				
$M_{SA,0}$			$\sigma^2_\omega > 0$	0.01	$\sigma^2_j \neq \sigma^2_\epsilon$	0.68
$M_{SA,J}$			$\sigma^2_\omega > 0$	0.01		
$M_{SPAP,0}$					$\sigma^2_j \neq \sigma^2_\epsilon$	0.63
$M_{SPAP,J}$						

expected if there were no such effects. However, $M_{S,0}$ is also inadequate. While there is *at least* marginal evidence of animal effects (recall that the values in Table VI are upper bounds), there is a very strong indication that species*paradigm interaction effects are also needed. The model with species and species*paradigm effects does not appear to need an animal effect, while the model with species and animal effects does need species*paradigm effects. Thus, the preferred model is $M_{SP,0}$. Not only do none of the models appear to need assessment-specific error variances, but whether or not this assumption is relaxed, the conclusions about which random effects need to be in the model are unaltered. Obviously there are other models that could be considered, and even different discrepancy measures that could be used on the models that were considered. Nevertheless, the sequences of models we tested and fitted suggest that model $M_{SP,0}$ appears adequate.

While the evidence supporting model $M_{SP,0}$ might seem compelling, there are several reasons to temper any conclusions from these results alone. One potential drawback of the discrepancy measures used is that they might not be particularly adept at detecting model violations. Furthermore, it is possible that the upper bounds are much too conservative. And perhaps there are simply not enough data to detect relatively small effects. In order to alleviate any concerns that might exist, posterior summaries on variance parameters are provided for all models in Table VII. The omission of animal and animal*paradigm interaction effects seems justified as they are relatively minor when included; together they explain on average less than 4% of the total variability. Figure 2 displays gaussian-kernel density estimates for the marginal posterior density of σ^2_γ and of σ^2_η under $M_{SPAP,0}$. The kernel density estimates are only graphed for nonnegative values of the variance. The actual estimates incorrectly suggest that the posterior densities do not pass through the origin and that negative values lie in the support. Regardless, the densities are concentrated near 0, again

Table VII. Variance parameter summaries for all models considered.

Model	σ^2_ϵ	σ^2_θ	σ^2_γ	σ^2_ω	σ^2_η
$M_{0,0}$	1.00 (—)				
$M_{S,0}$	0.88 (0.07)	0.12 (0.07)			
$M_{S,J}$	0.87 (0.07)	0.13 (0.07)			
$M_{SP,0}$	0.68 (0.07)	0.15 (0.08)		0.18 (0.05)	
$M_{SP,J}$	0.68 (0.08)	0.15 (0.08)		0.17 (0.05)	
$M_{SA,0}$	0.88 (0.06)	0.11 (0.06)	0.012 (0.007)		
$M_{SA,J}$	0.85 (0.07)	0.13 (0.07)	0.016 (0.009)		
$M_{SPAP,0}$	0.67 (0.07)	0.13 (0.07)	0.0050 (0.004)	0.17 (0.05)	0.022 (0.01)
$M_{SPAP,J}$	0.68 (0.07)	0.13 (0.07)	0.0056 (0.005)	0.16 (0.05)	0.030 (0.01)

confirming their relative unimportance when included in the model.

C. Results from Selected Model

The preferred model has species effects, species*paradigm effects, and assumes constant error variance across assessments. The model fit is now described in more detail, including comments on mixing and convergence of the Markov chain as well as posterior summaries.

Some attention was given to examining convergence and mixing in the preliminary MCMC runs of various models under consideration, but the convergence and mixing should be more rigorously examined for any models on which inferences are to be made. In addition to the preliminary run of the selected model, two much longer runs of the MCMC algorithm were used with different starting points. For each of

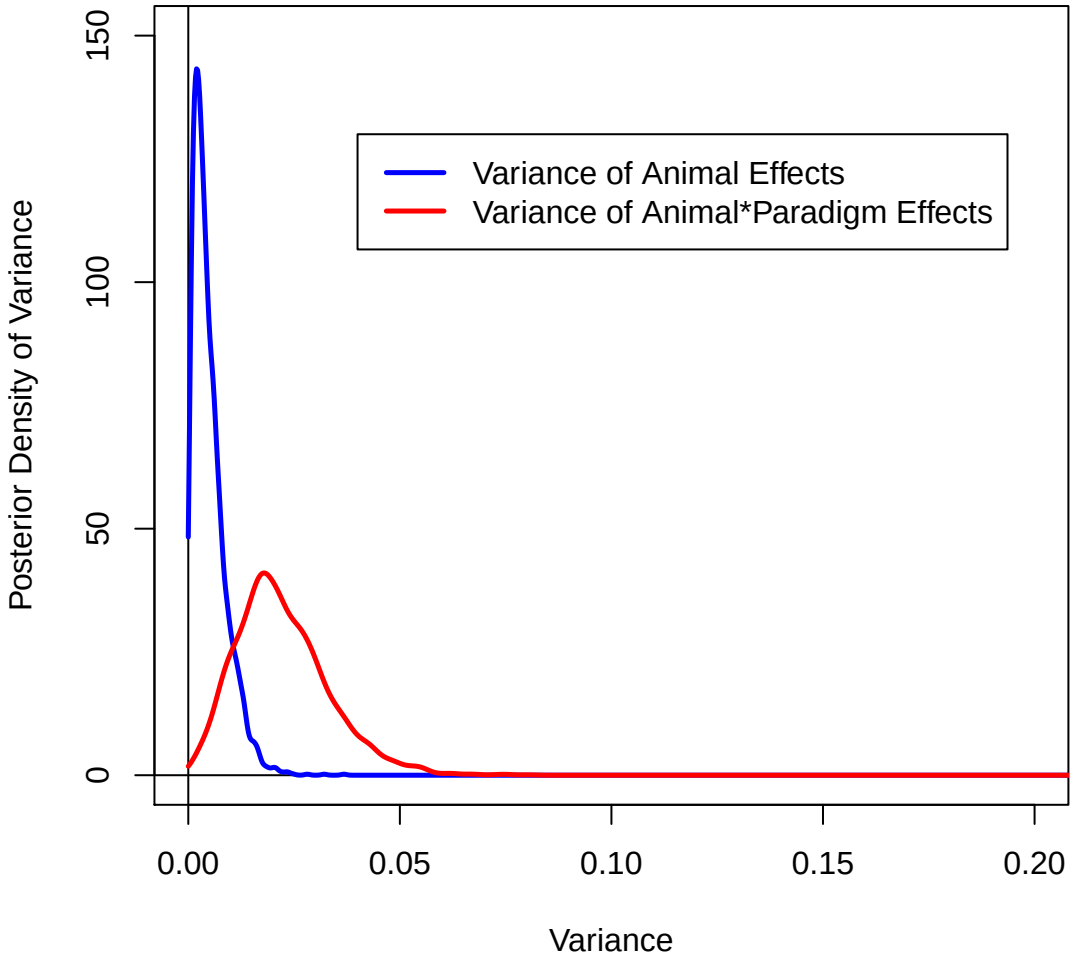

Fig. 2. Kernel density estimates of the marginal posterior distributions for σ^2_γ and σ^2_η under $M_{SPAP,0}$. The kernel density estimation does not enforce the constraints that the posterior density of σ^2_γ and of σ^2_η at 0 must be 0, but the concentration of both densities near zero is nonetheless apparent, particularly for the animal effects.

these runs, there was a lengthy burn-in period (200,000 or 250,000 iterations) and an additional large number of iterations (6,000,000 or 7,500,000). The latter run is used for final inferences, but the former lengthy run suggests good agreement between the two runs and thus suggests that the runs have converged to the target (posterior) distribution. All iterations after the burn-in period were used for estimating the posterior mean and posterior standard deviation of selected model parameters. However, because of computer memory limitations, the sequences were thinned by using every 500^{th} iteration for most graphical displays, for 95% central credible intervals, and for the posterior mean and posterior standard deviation of combined species and species*paradigm effects.

Appendix A contains trace plots of various parameters and information on tuning parameters to get the target acceptance rates. While this is important information for practical purposes and to justify validity of the subsequent model inferences, the detailed information is relegated to the Appendix because it is not of principal interest. It is important to note, however, that the chain appears to converge rather quickly although it mixes poorly. The thinned chain appears to mix well, but it is unsettling that the chain must be drastically thinned in order to appear to mix well. The large number of iterations offsets somewhat the effects of thinning, but it also implies a greater computational burden.

This model has random effects for both species and species*paradigm effects, so naturally one of the first questions of interest is the relative importance of each. Recall that the model has been constructed so that the prior predictive distribution of a latent variable z_{ijt} has a variance of exactly one by requiring that $\sigma_\epsilon^2 + \sigma_\theta^2 + \sigma_\omega^2 = 1$. Thus, each variance parameter in $\boldsymbol{\sigma}$ can be interpreted as the proportion of total variability in a predictive distribution (for a new animal and species) that is explained by the corresponding effect. Figure 3 displays several contours of the

posterior distribution of $\boldsymbol{\sigma}$. While σ_ϵ^2 is not depicted, it is implicit because of the prior constraint.

Posterior inference on $\boldsymbol{\sigma}$ suggests that the species and species*paradigm effects have fairly similar variability. However, the error variability tends to be the primary source of variability in $\mathbf{z}$, accounting for more than two-thirds of the total variability on average. Table VIII contains posterior summaries of the the model's nuisance parameters, $\boldsymbol{\kappa}$ and $\boldsymbol{\tau}$. It is interesting to note that the prior density for $\boldsymbol{\tau}$ (each τ_j iid $Cauchy(0, 0.5^2)$) seems to have been inappropriate for τ_5 and τ_{11}. Section D contains results from a sensitivity analysis which uses less informative priors on $\boldsymbol{\kappa}$ and $\boldsymbol{\tau}$.

Besides the variance parameters, other model quantities of note are the random effects. Rather than interpret them individually, we focus on the combined effects (i.e., $\theta_{s(i)}+\omega_{s(i),g(j)}$) for each paradigm. Table IX and Table X contain numerical summaries of these combined effects, while Figure 4 graphically depicts the estimated marginal posterior distributions. There is dramatic variation in the standard deviations of these combined effects across paradigms. This is not surprising because both the number of trials in binomial assessments and the number of assessments per paradigm varied substantially. Furthermore, for the last two paradigms none of the apes were observed, and the gaze following paradigm also did not include any observations from long-tailed macaques.

Figure 5 has six image plots. Each plot focuses on one of the paradigms. The plot regions are shaded according to the probability that the species in the row has a larger combined effect $\theta + \omega$ than the species in the column. The color of each grid box is determined by the posterior probability that the combined effect for the species listed in the grid's *row* is larger than that of the the species in the grid's *column*. The colors range from dark blue to white and then to dark red as the probabilities change from 0 to 0.5 and then to 1. A dark blue box indicates a posterior probability

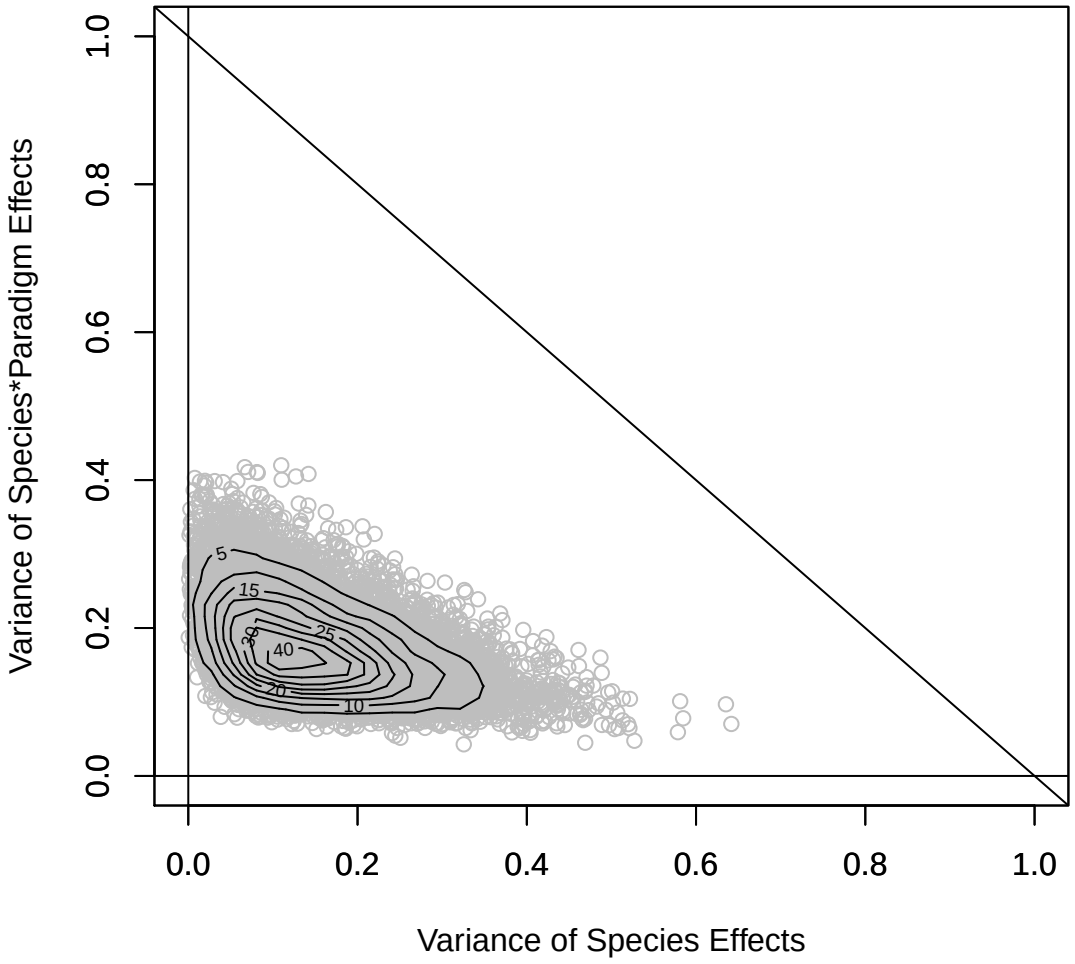

Fig. 3. Posterior distribution of the variance parameters σ_θ^2 (for species effects), σ_ω^2 (for species*paradigm effects), and $\sigma_\epsilon^2 \equiv 1 - \sigma_\theta^2 - \sigma_\omega^2$ (for error terms). The posterior density is defined as zero outside of the demarcated triangular region.

Table VIII. Posterior summaries of model nuisance parameters.

Parameter	Posterior Mean	Posterior SD	95% Credible Interval
κ_1	1.79	0.91	(0.63, 4.04)
κ_2	1.46	0.73	(0.54, 3.33)
τ_3	-0.23	0.21	(-0.64, 0.16)
κ_4	0.035	0.010	(0.019, 0.056)
τ_5	1.66	0.21	(1.23, 2.05)
τ_6	-0.43	0.20	(-0.83, -0.05)
τ_7	-0.06	0.19	(-0.45, 0.30)
τ_8	-0.23	0.19	(-0.61, 0.14)
τ_9	-0.10	0.19	(-0.48, 0.26)
τ_{10}	-0.27	0.21	(-0.71, 0.12)
τ_{11}	-0.99	0.27	(-1.53, -0.48)
τ_{12}	-0.18	0.14	(-0.47, 0.09)
τ_{13}	0.05	0.14	(-0.23, 0.32)
τ_{14}	-0.16	0.14	(-0.44, 0.12)
τ_{15}	-0.12	0.14	(-0.40, 0.15)
τ_{16}	0.05	0.14	(-0.22, 0.32)
τ_{17}	-0.01	0.14	(-0.29, 0.26)
κ_{18}	0.22	0.10	(0.09, 0.46)
κ_{19}	1.97	1.06	(0.62, 4.69)
κ_{20}	0.015	0.007	(0.005, 0.033)

Table IX. Posterior summaries from the first three paradigms of the combined species and species*paradigm random effects. These summaries are based on the thinned sequences after burn-in. The 95% credible intervals are equal-tailed.

		Posterior Quantity		
Paradigm	Species	Mean	SD	95% Credible Interval
Inhibition	Chimpanzee	0.57	0.22	(0.13, 0.98)
	Bonobo	0.47	0.24	(-0.01, 0.92)
	Gorilla	0.05	0.23	(-0.41, 0.48)
	Orangutan	1.05	0.21	(0.62, 1.46)
	Spider monkey	0.03	0.21	(-0.39, 0.42)
	Capuchin monkey	-0.31	0.21	(-0.73, 0.09)
	Long-tailed macaque	-0.70	0.23	(-1.16, -0.27)
Memory	Chimpanzee	0.77	0.24	(0.30, 1.23)
	Bonobo	0.88	0.28	(0.34, 1.43)
	Gorilla	0.18	0.23	(-0.28, 0.64)
	Orangutan	0.01	0.22	(-0.43, 0.43)
	Spider monkey	0.30	0.21	(-0.10, 0.71)
	Capuchin monkey	-0.31	0.21	(-0.73, 0.09)
	Long-tailed macaque	-0.30	0.21	(-0.72, 0.11)
Transposition	Chimpanzee	0.85	0.25	(0.37, 1.34)
	Bonobo	1.10	0.33	(0.49, 1.79)
	Gorilla	0.41	0.23	(-0.05, 0.85)
	Orangutan	0.29	0.23	(-0.17, 0.75)
	Spider monkey	-0.20	0.20	(-0.61, 0.19)
	Capuchin monkey	-0.45	0.21	(-0.87, -0.05)
	Long-tailed macaque	-0.02	0.20	(-0.43, 0.37)

Table X. Posterior summaries from the last three paradigms of the combined species and species*paradigm random effects. These summaries are based on the thinned sequences after burn-in. The 95% credible intervals are equal-tailed.

		Posterior Quantity		
Paradigm	Species	Mean	SD	95% Credible Interval
Support	Chimpanzee	0.20	0.14	(-0.07, 0.46)
	Bonobo	0.23	0.15	(-0.06, 0.52)
	Gorilla	0.09	0.15	(-0.21, 0.38)
	Orangutan	0.19	0.15	(-0.11, 0.48)
	Spider monkey	0.61	0.14	(0.33, 0.88)
	Capuchin monkey	0.24	0.14	(-0.04, 0.50)
	Long-tailed macaque	-0.02	0.14	(-0.30, 0.25)
Gaze following	Chimpanzee	0.43	0.48	(-0.54, 1.35)
	Bonobo	0.49	0.48	(-0.50, 1.41)
	Gorilla	0.14	0.46	(-0.77, 1.04)
	Orangutan	0.28	0.47	(-0.66, 1.19)
	Spider monkey	-0.09	0.35	(-0.80, 0.60)
	Capuchin monkey	0.09	0.35	(-0.61, 0.79)
	Long-tailed macaque	-0.19	0.46	(-1.07, 0.72)
Reversed	Chimpanzee	0.44	0.48	(-0.52, 1.36)
contingency	Bonobo	0.49	0.48	(-0.48, 1.43)
	Gorilla	0.14	0.46	(-0.79, 1.05)
	Orangutan	0.28	0.47	(-0.66, 1.17)
	Spider monkey	-0.36	0.33	(-1.01, 0.28)
	Capuchin monkey	0.41	0.33	(-0.23, 1.07)
	Long-tailed macaque	-0.22	0.33	(-0.85, 0.42)

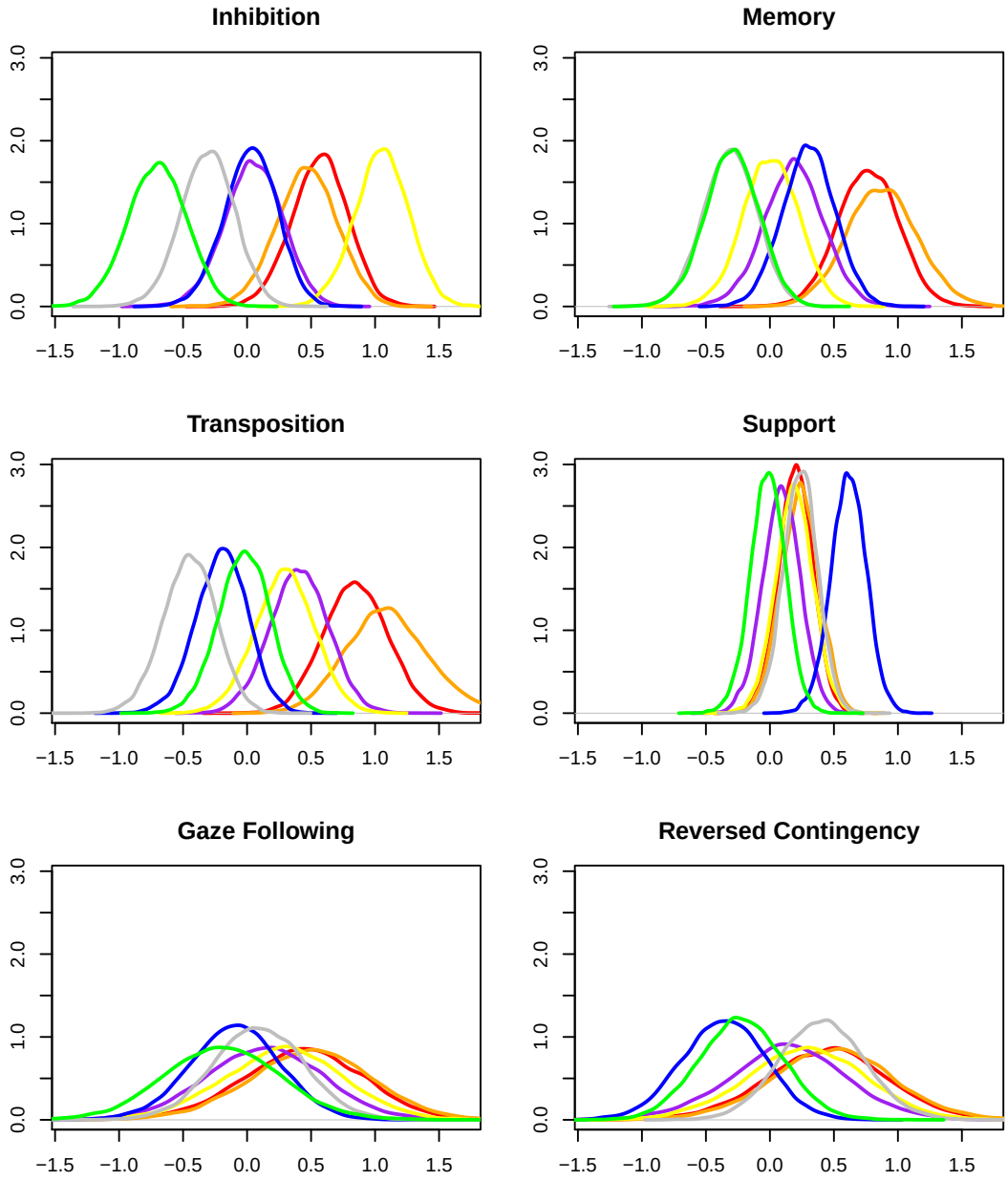

Fig. 4. Kernel density estimates of the marginal posterior distributions for $\theta_{s(i)} + \omega_{s(i),g(j)}$ by paradigm. The line color/species associations are red—chimpanzees, orange—bonobos, purple—gorillas, yellow—orangutans, blue—spider monkeys, gray—capuchin monkeys, and green—long-tailed macaques.

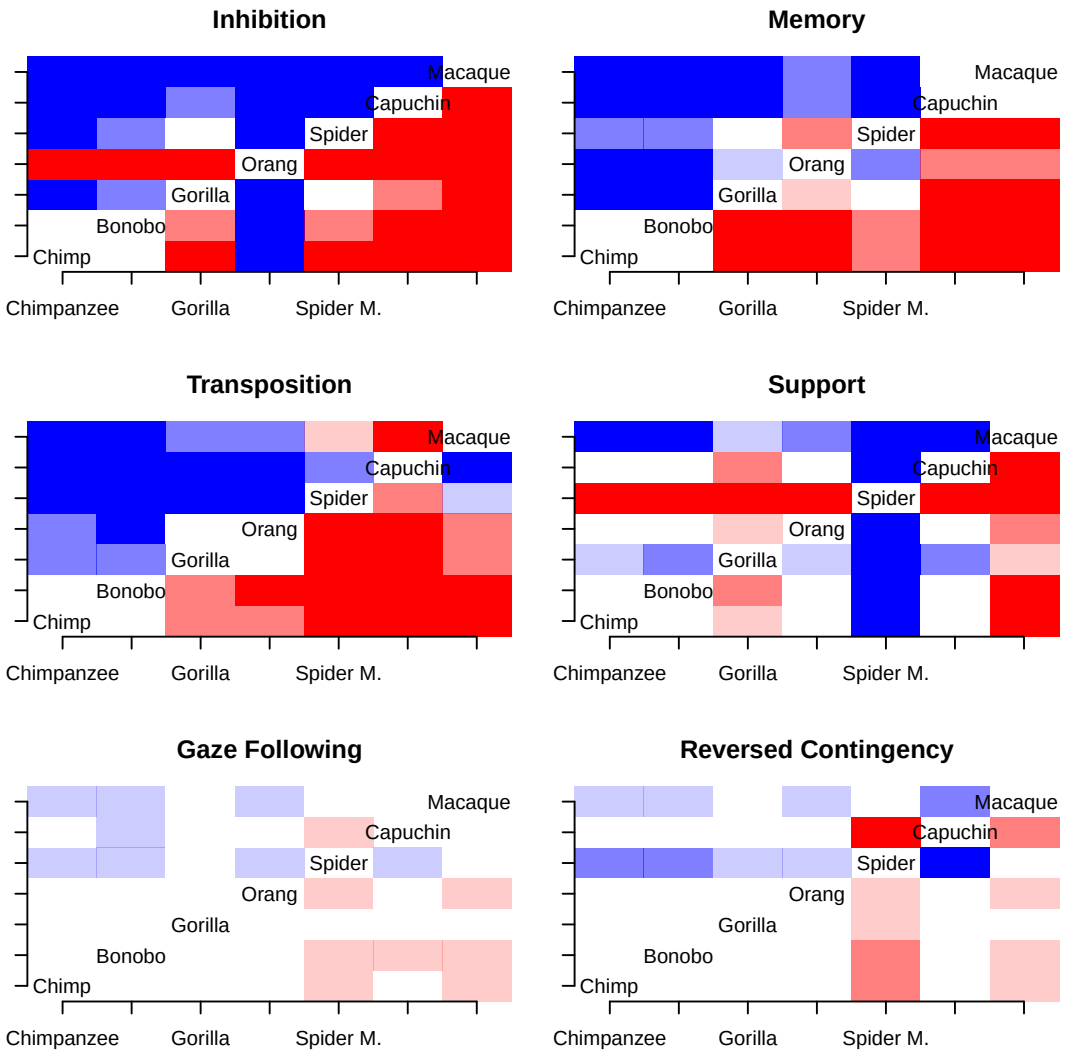

Fig. 5. Image plots of the posterior probability that the listed row species is better than the listed column species for the specified paradigm. A dark blue box indicates a posterior probability of at most 0.01, a light blue indicates posterior probability in (0.01, 0.10], a faint blue indicates (0.10,0.25], a white indicates (0.25,0.75), a faint red indicates [0.75,0.90), a light red indicates [0.90,0.99), and a dark red indicates a posterior probability of at least 0.99. The diagonal boxes are artificially white to divert attention away from them because the comparison is meaningless (i.e., comparing each species with itself).

of at most 0.01. A light blue box indicates a posterior probability in (0.01, 0.10]. A faint blue box indicates a posterior probability in (0.10,0.25]. A white box indicates a posterior probability in (0.25,0.75) or the meaningless comparison of a species with itself. A faint red box indicates a posterior probability in [0.75,0.90). A light red box indicates a posterior probability in [0.90,0.99). Finally, a dark red box indicates a posterior probability of at least 0.99 that the listed row species has a larger combined effect than the species in the column. Note that the diagonal boxes are artificially white (rather than dark blue) to divert attention away from them because comparing each species with itself is completely uninformative.

For each plot in Figure 5, a row with only dark red (dark blue) boxes off the diagonal suggests strong evidence that the species listed in that row is better (worse) on average than the other species for that paradigm. Note that these comparisons cannot be performed using only the summary measures in Table IX and Table X, owing to the correlation in posterior draws of the combined effects.

Several aspects of Figure 5 suggest why the species*paradigm interactions seem to be so important for the model. First, orangutan performance in the inhibition paradigm (paradigm 1) is much better relative to the other species than might have been expected based on any other paradigm. Similarly, spider monkeys excelled relative to the other species in the support paradigm (paradigm 4) but were generally average in the other paradigms.

Other notable findings include the similarity of chimpanzees and bonobos in each of the six paradigms. Only the transposition paradigm (paradigm 3) is even mildly suggestive of a difference between the paradigm-specific intelligence of these species. Also, the great apes as a group performed better than the monkeys in the first three paradigms. Other than exceptionally good (poor) performances by spider monkeys (long-tailed macaques), there were not great differences for paradigm 4.

The final two paradigms must be interpreted with caution; while there is not strong posterior evidence of differences between most pairs of species, this might reasonably be attributed to two factors. First, no great apes were observed for any assessments from either of these paradigms, and paradigm 5 only used observations from two monkey species. Second, the paradigms consisted of either (a) only two assessments, each with only two trials involved in the binomial response; or (b) only one assessment, having rank response. Therefore, it not surprising that species comparisons for these paradigms are rather inconclusive.

D. Model Sensitivity

In order to assess the robustness of these findings, results from several alternative modeling strategies are now considered. The changes may be summarized as either: (a) using somewhat more vague priors on the nuisance parameters $\boldsymbol{\tau}$ and $\boldsymbol{\kappa}$; or (b) excluding the last two paradigms.

In order to examine the effect of the selected hyperparameter values, the selected model was refitted after making some modest changes to have less informative priors. Specifically the priors were changed as follows.

- $\tau_j \sim Cauchy(0, 1^2)$ instead of $\tau_j \sim Cauchy(0, 0.5^2)$
- $\kappa_j \sim Gamma(0.5, 0.5)$ instead of $\kappa_j \sim Gamma(1, 1)$

Also, because the observed data for the fifth and sixth paradigms are limited to only two or three monkey species and one or two assessments, these paradigms were excluded to gauge their effect on model fit.

Posterior inference is affected somewhat by the new prior specification and with the exclusion of the last three assessments. Table XI reports the posterior mean and

standard deviation for several quantities of interest for $M_{SP,0}$ with the original analysis, alternate prior, and omission of the last two paradigms. The alternate prior lead to a downward shift of roughly 0.1 in the species effects and the threshold parameters. This change is not surprising because 13 of the 14 binomial assessments (and the majority of the total number of observed binomial trials) had a negative posterior mean for the assessment's threshold. Increasing the scale of the prior distribution on $\boldsymbol{\tau}$ diminishes the shrinkage of these parameters to zero, and since nearly all of the τ_j's were negative the cumulative effect of the species and species*paradigm effects also drops. What is somewhat surprising is that this change seems to have affected the species effects more than the species*paradigm effects. The alternate prior also has values of κ that tend to be further away from one (the prior mean) than the values were in the original analysis.

Omitting the last two paradigms (assessments 18–20) generally had very little effect compared to the original modeling with data from all six paradigms. The primary difference is that the species effects for spider monkeys (species 5) and capuchin monkeys (species 6) changed somewhat. That they changed is not surprising because these are the only two species that were represented in the gaze-following paradigm (paradigm 5) and they are two of the three species represented in the reversed-contingency paradigm (paradigm 6). The species*paradigm interaction effects have adjusted by roughly the opposite amount as the species adjustments for spider monkeys and capuchin monkeys, indicating that the combined species and species*paradigm effects ($\theta_{s(i)}+\omega_{s(i),g(j)}$) for the first four paradigms have not changed much by excluding the last two paradigms.

Figure 6 and Figure 7 can be compared with Figure 4 to understand the sensitivity of the combined random effects to the prior and the last two paradigms. Figure 8 and Figure 9 can be compared with Figure 5 to gauge the impact of such changes on

Table XI. Posterior summaries under original model and variations.

	Posterior Mean (SD)		
Model	Original	Alternate	Omit 5^{th}, 6^{th}
Quantity	Analysis	Prior	Paradigms
σ^2_ϵ	0.68 (0.08)	0.70 (0.07)	0.67 (0.08)
σ^2_θ	0.15 (0.08)	0.12 (0.07)	0.17 (0.09)
σ^2_ω	0.17 (0.05)	0.18 (0.05)	0.16 (0.05)
θ_1	0.44 (0.23)	0.34 (0.22)	0.46 (0.22)
θ_2	0.49 (0.24)	0.39 (0.23)	0.52 (0.23)
θ_3	0.14 (0.20)	0.06 (0.20)	0.15 (0.20)
θ_4	0.28 (0.21)	0.19 (0.21)	0.30 (0.21)
θ_5	0.04 (0.18)	-0.04 (0.19)	0.16 (0.20)
θ_6	-0.04 (0.18)	-0.12 (0.19)	-0.15 (0.20)
θ_7	-0.19 (0.19)	-0.26 (0.20)	-0.19 (0.20)
$\omega_{1,1}$	0.13 (0.26)	0.14 (0.26)	0.12 (0.25)
$\omega_{2,1}$	-0.02 (0.27)	-0.01 (0.27)	-0.03 (0.26)
$\omega_{3,1}$	-0.09 (0.25)	-0.10 (0.25)	-0.08 (0.24)
$\omega_{4,1}$	0.77 (0.25)	0.78 (0.25)	0.76 (0.24)
$\omega_{5,1}$	-0.01 (0.23)	-0.03 (0.23)	-0.11 (0.23)
$\omega_{6,1}$	-0.27 (0.23)	-0.29 (0.24)	-0.15 (0.24)
$\omega_{7,1}$	-0.52 (0.25)	-0.55 (0.26)	-0.50 (0.25)
κ_1	1.79 (0.91)	2.11 (1.26)	1.78 (0.91)
κ_2	1.46 (0.73)	1.63 (0.94)	1.46 (0.73)
κ_4	0.035 (0.010)	0.034 (0.009)	0.035 (0.010)
τ_3	-0.23 (0.21)	-0.33 (0.22)	-0.21 (0.21)
τ_5	1.66 (0.21)	1.59 (0.22)	1.67 (0.21)
τ_6	-0.43 (0.20)	-0.55 (0.21)	-0.42 (0.20)

interspecies comparisons for each paradigm.

E. Discussion of Results

The combined primate intelligence data set was very exciting because of its ability to address questions regarding the necessity and relative importance of species, species*paradigm, animal, and animal*paradigm random effects on non-human primate assessments from different paradigms of intelligence testing. Several key findings in the data analysis are that there are substantial species main effects and similarly important species*paradigm interaction effects. That is, intelligence seems to systematically differ across species of great apes and monkeys, but a reduction to a general intelligence latent trait is not adequate.

Johnson et al. (2002) did not find strong evidence of genus*paradigm interactions in their meta-analysis of primate intelligence tests (though genus effects were included). While it is difficult to identify the exact nature of this seemingly contradictory finding, there are several possible explanations. Among these are differences in assessments, paradigms, and species across the studies. Another possible reason is that the current data set contains mostly assessments using binomial responses, and with several of them having six or more trials there is arguably more information than if all of the binomial outcomes had been converted to rank responses. Perhaps in this sense the current study is better suited to pick up effects.

Yet another explanation is that there are key differences in interpretation of random effects between the proposed methodology and the rank data methodology introduced by Johnson et al. (2002). In particular, consider a model where the latent variable is the sum of a random species effect and an error term. The rank methodology of the aforementioned authors would establish scale by constraining the variance

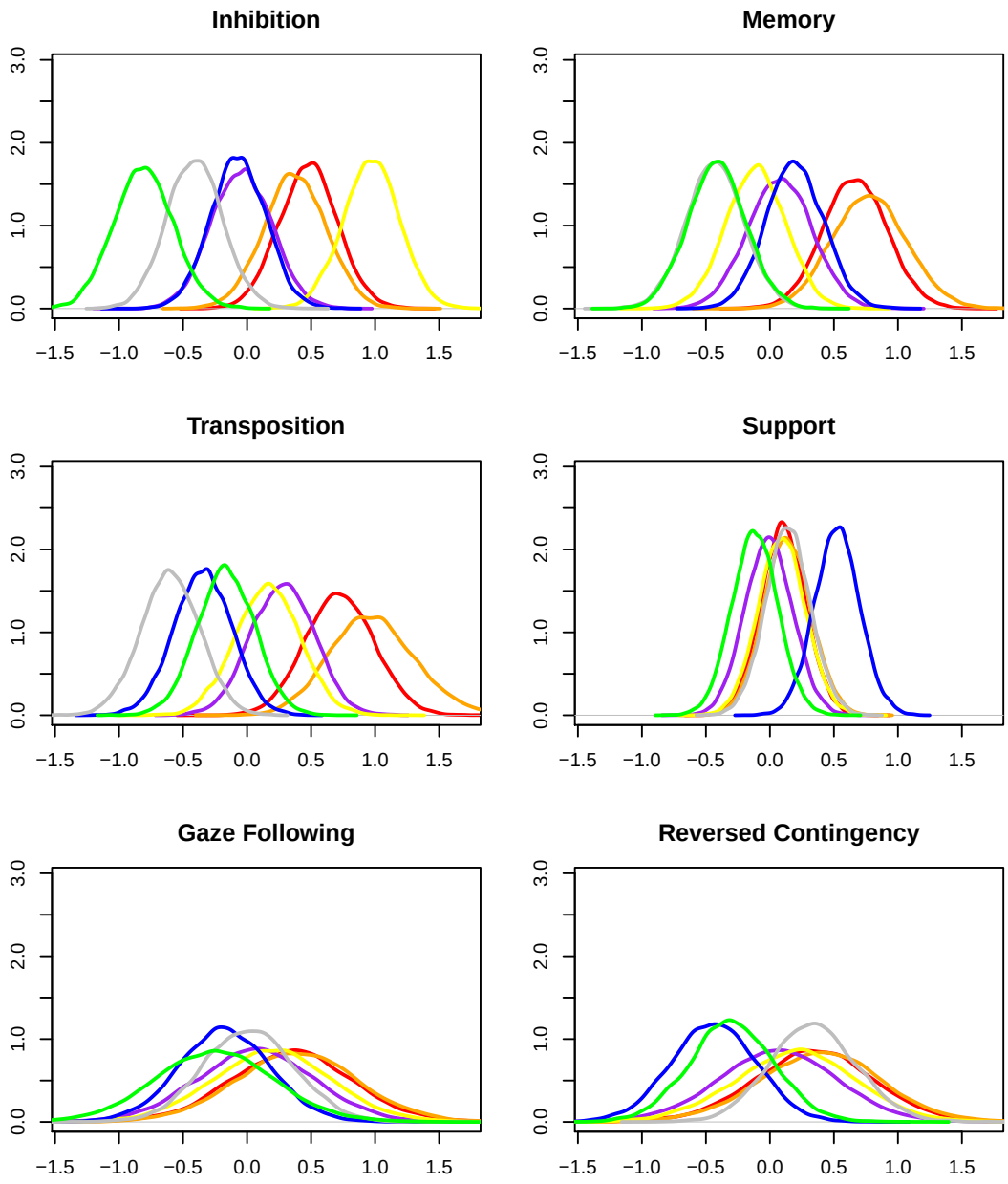

Fig. 6. Kernel density estimates of the marginal posterior distributions for $\theta_{s(i)} + \omega_{s(i),g(j)}$ by paradigm *when the prior distribution is less informative*. The line color/species associations are red—chimpanzees, orange—bonobos, purple—gorillas, yellow—orangutans, blue—spider monkeys, gray—capuchin monkeys, and green—long-tailed macaques.

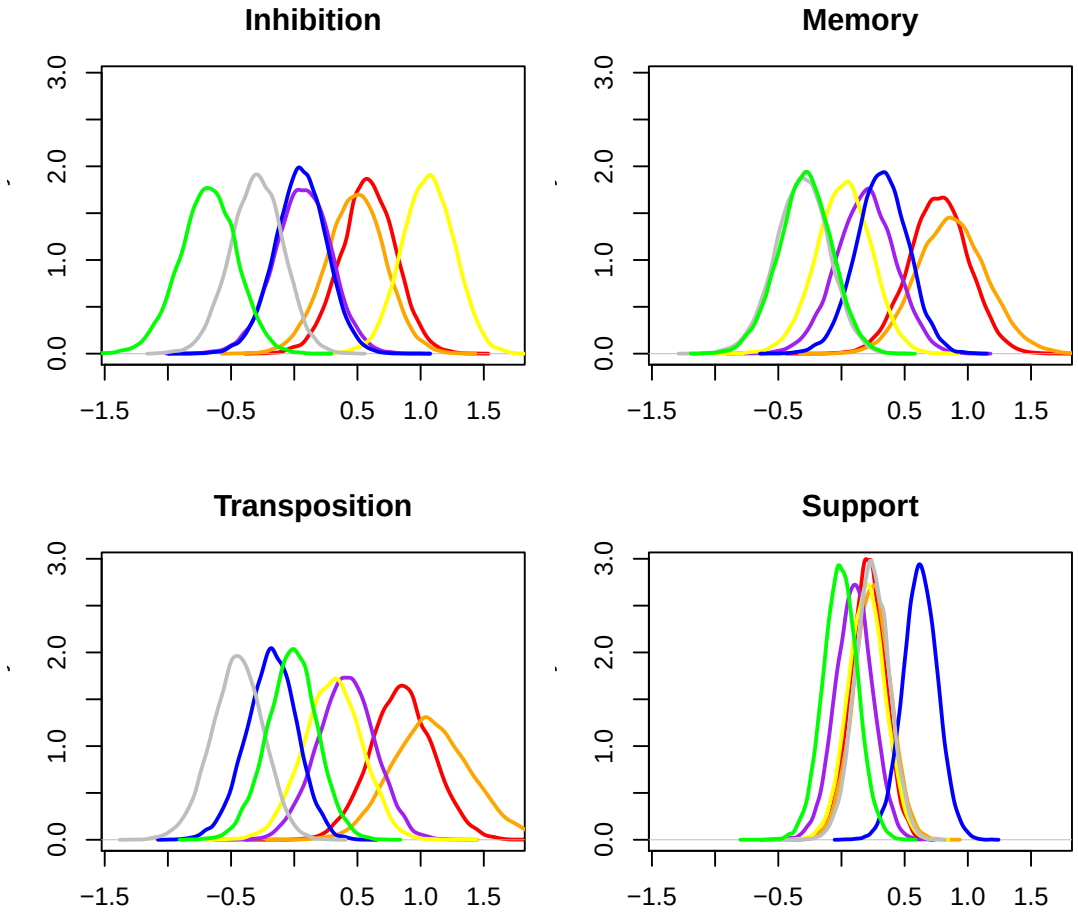

Fig. 7. Kernel density estimates of the marginal posterior distributions for $\theta_{s(i)}+\omega_{s(i),g(j)}$ by paradigm *when data from the last two paradigms are excluded*. The line color/species associations are red—chimpanzees, orange—bonobos, purple—gorillas, yellow—orangutans, blue—spider monkeys, gray—capuchin monkeys, and green—long-tailed macaques.

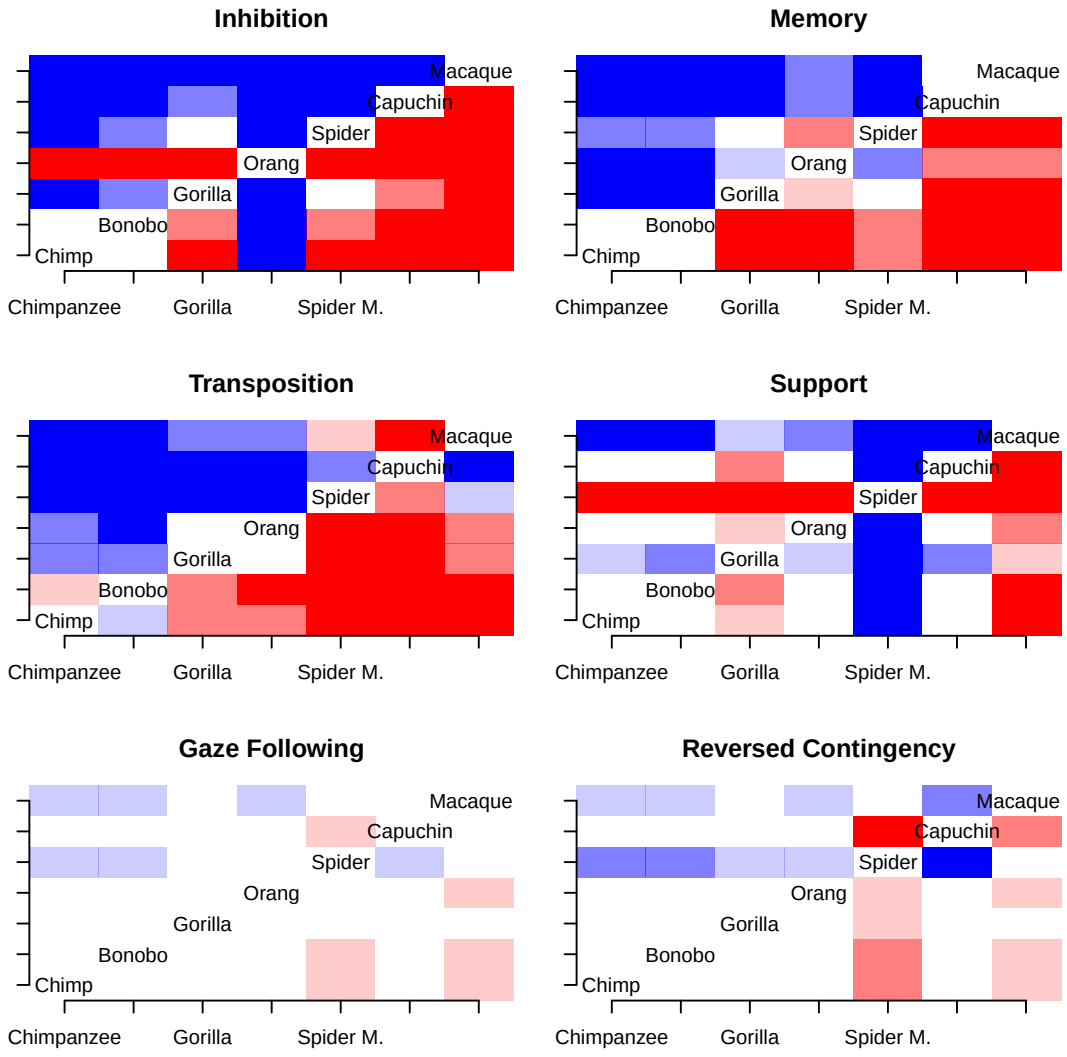

Fig. 8. Image plots of the posterior probability that the listed row species is better than the listed column species for the specified paradigm *when the prior distribution is less informative*. A dark blue box indicates a posterior probability of at most 0.01, a light blue indicates posterior probability in (0.01, 0.10], a faint blue indicates (0.10,0.25], a white indicates (0.25,0.75), a faint red indicates [0.75,0.90), a light red indicates [0.90,0.99), and a dark red indicates a posterior probability of at least 0.99. The diagonal boxes are artificially white to divert attention away from them because the comparison is meaningless (i.e., comparing each species with itself).

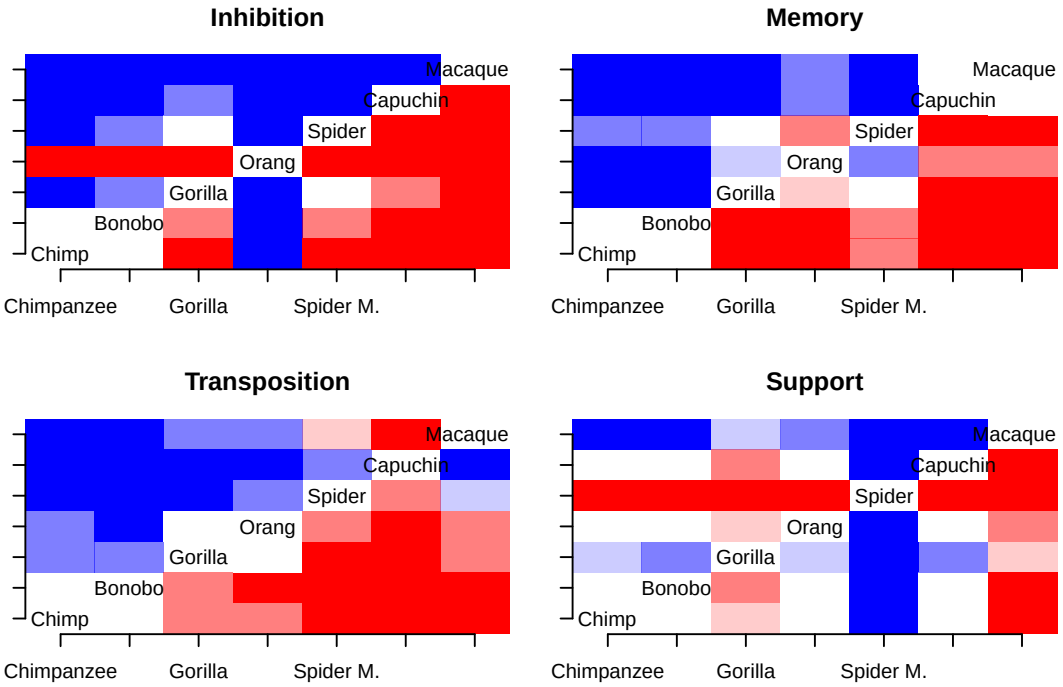

Fig. 9. Image plots of the posterior probability that the listed row species is better than the listed column species for the specified paradigm *when data from the last two paradigms are excluded*. A dark blue box indicates a posterior probability of at most 0.01, a light blue indicates posterior probability in (0.01, 0.10], a faint blue indicates (0.10,0.25], a white indicates (0.25,0.75), a faint red indicates [0.75,0.90), a light red indicates [0.90,0.99), and a dark red indicates a posterior probability of at least 0.99. The diagonal boxes are artificially white to divert attention away from them because the comparison is meaningless (i.e., comparing each species with itself).

of one of the species random effects to equal one and would let each assessment have its own error variance; thus, $Var(z_{ijt}|\sigma_j^2, \sigma_\theta^2 \equiv 1) = 1 + \sigma_j^2$. As σ_j^2 increases, the relative importance of the species effects diminishes. If error variances differ systematically across paradigms, this implies different relative contributions of the species effect across paradigms, and in a sense implies a species*paradigm interaction. On the other hand, the current methodology establishes scale in the latent variables $\mathbf{z}$ by constraining their prior marginal variance to equal 1, regardless of the number of random effects involved. When each error variance is constrained to equal σ_ϵ^2, as in $M_{S,0}$, the relative contribution of the random species effect is assumed to be the same across paradigms. And even when the error variance is allowed to differ across assessments, as in $M_{S,J}$, the σ_j^2 parameters are modeled hierarchically around σ_ϵ^2 with informative priors and thus tend to be close together.

If species*paradigm interactions are important for distinguishing the degree of separation between species for each paradigm but do not change the ordering, the aforementioned methodology might not be able to pick this up, while model $M_{S,0}$ would be better able and even $M_{S,J}$ might be able to do so. However, this explanation seems unsatisfactory because the present model had little evidence to support assessment-specific error variances and more importantly because the ordering of the combined species and species*paradigm effects changed, sometimes dramatically, across paradigms.

Whatever the correct explanation may be, the finding is that intelligence is paradigm specific, or at least that different paradigms tend to yield different conclusions regarding interspecies intelligence comparisons. This is consistent with conclusions by, for example, Riopelle and Hill (1973, p. 541).

There is little evidence from the current study to support the presence of considerable animal or animal*paradigm effects unless species*paradigm effects are ignored.

Together, the animal and animal*paradigm effects explained on average less than 4% of the total variability in the latent variables. On the other hand, the species and species*paradigm effects together explained roughly ten times as much of the variability as the animal-related effects. This suggests that there is substantially more systematic interspecies variation than systematic intraspecies variation, at least for the types of animals tested. This finding may not hold in studies with greater variation in animal age, different assessments and paradigms, or different species. Care must be taken in interpreting the results.

CHAPTER V

CONCLUSIONS AND POSSIBLE EXTENSIONS

Both binomial response and rank response data can be analyzed using latent variables to form augmented data. We used the augmentation approach to jointly model binomial and rank data using a random effects model for the latent variables. We also used Dirichlet priors to establish scale for the latent variables and readily admit non-local priors for random effects. We detailed 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. Finally, we demonstrated model comparison using discrepancy measures based on pivotal quantities because estimation of Bayes factors can be especially challenging in large data sets with many latent variables and non-local priors.

The primate intelligence application helped to answer several questions and demonstrated the usefulness of the methodology. In particular, the discrepancy measures appeared to effectively detect needed random effects while not flagging effects that were not strongly supported by the data, suggesting both sensitivity and specificity.

We extended the model to allow for assessment-specific error variances using a hierarchical model. Further work could relax the assumption that all levels of a given random effect have the same variability. In addition, future work could involve a more general and user-friendly computer program to implement posterior sampling and calculation of discrepancy measures. We wrote a custom program in C++ using the GNU Scientific Library (Galassi et al. 2009) to fit the primate intelligence models and compute discrepancy measures, but the program still requires more user modifi-

cation than desirable for general use. Because R (R Development Core Team 2010) was used to provide input data to the C++ program and for summarizing the output information, it would be ideal to provide an R wrapper that would make fitting these types of models and obtaining useful summaries even more automatic for practitioners. However, creating such an R wrapper is challenging because of memory allocation issues. Meanwhile, a direct R implementation of the MCMC algorithm would likely be too slow to be of practical use. For the primate intelligence application, the C++ implementation only obtained about 200K iterations/hour and yet an attempt to use R was about two orders of magnitude slower.

Another possible avenue for future work is to allow continuous effects and fixed factor effects in the hierarchical modeling of latent variables. In particular, age and sex might be important in modeling primate intelligence data. The inclusion of these explanatory variables would require a fundamental change in the modeling approach to establish scale of the latent variables.

Finally, other methods of approximating Bayes factors could be explored. Discrepancy measures can be very useful, as demonstrated by the analysis of Chapter IV, but there are reasons to prefer the Bayes factor in model selection techniques. One of these reasons is that the discrepancy measures based on pivotal quantities ignored some data to avoid a poor chi-square approximation when expected counts were low. Another is that while the discrepancy measures were used to determine a bound on the PPP p-value, having a bound is, in general, certainly less desirable than having the actual PPP p-value.

The techniques of the previous chapters and their application to primate intelligence testing represent an important step forward even without the possible extensions just stated. 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.

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., Aureli, F., and Call, J. (2008), "Fission-Fusion Dynamics, Behavioral Flexibility, and Inhibitory Control in Primates," *Current Biology*, 18, 1415–1419.

— (2010), "Monkeys and Apes: Are Their Cognitive Skills Really So Different?" *American Journal of Physical Anthropology*, 143, 188–197.

Amici, F., Aureli, F., Visalberghi, E., and Call, J. (2009), "Spider Monkeys (*Ateles geoffroyi*) and Capuchin Monkeys (*Cebus apella*) Follow Gaze Around Barriers: Evidence for Perspective Taking?" *Journal of Comparative Psychology*, 123, 368–374.

Barth, J. and Call, J. (2006), "Tracking the Displacement of Objects: A Series of Tasks With Great Apes (*Pan troglodytes*, *Pan paniscus*, *Gorilla gorilla*, and *Pongo pygmaeus*) and Young Children (*Homo sapiens*)," *Journal of Experimental Psychology*, 32, 239–252.

Chib, S. (1995), "Marginal Likelihood from the Gibbs Output," *Journal of the American Statistical Association*, 90, 1313–1321.

— (2001), "Markov Chain Monte Carlo Methods: Computation and Inference," in *Handbook of Econometrics*, Vol. 5, eds. J. J. Heckman and E. Leamer, Amsterdam: Elsevier, pp. 3569–3649.

Chib, S. and Jeliazkov, I. (2001), "Marginal Likelihood from the Metropolis-Hastings Output," *Journal of the American Statistical Association*, 96, 270–281.

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

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

Galassi, M., Davies, J., Theiler, J., Gough, B., Jungman, G., Alken, P., Booth, M., and Rossi, F. (2009), *GNU Scientific Library Reference Manual* (3rd ed.), Godalming, U.K.: Network Theory Limited.

Gelfand, A. E. and Dey, D. K. (1994), "Bayesian Model Choice: Asymptotics and Exact Calculations," *Journal of the Royal Statistical Society,* Ser. B, 56, 501–514.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004), *Bayesian Data Analysis* (2nd ed.), Boca Raton, FL: Chapman & Hall/CRC.

Harlow, H. F. and Mears, C. (1979), *The Human Model: Primate Perspectives*, Washington, D.C.: V. H. Winston & Sons.

Herrmann, E., Wobber, V., and Call, J. (2008), "Great Apes' (*Pan troglodytes*, *Pan paniscus*, *Gorilla gorilla* , *Pongo pygmaeus*) Understanding of Tool Function Properties After Limited Experience," *Journal of Comparative Psychology*, 122, 220–230.

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

Johnson, V. E. and Albert, J. H. (1999), *Ordinal Data Modeling*, New York: Springer-Verlag.

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.

Kass, R. E. and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.

Lopes, H. F. and West, M. (2004), "Bayesian Model Assessment in Factor Analysis," *Statistica Sinica*, 14, 41–67.

Marden, J. I. (1995), *Analyzing and Modeling Rank Data*, London: Chapman & Hall.

Meng, X.-L. and Wong, W. H. (1996), "Simulating Ratios of Normalizing Constants via a Simple Identity: A Theoretical Exploration," *Statistica Sinica*, 6, 831–860.

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

R Development Core Team (2010), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, available at *http://www.R-project.org*.

Raftery, A. E. (1993), "Discussion on the Meeting on the Gibbs Sampler and Other Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society,* Ser. B, 55, 85.

Riopelle, A. J. and Hill, C. W. (1973), "Complex Processes," in *Comparative Psychology: A Modern Survey*, eds. D. A. Dewsbury and D. A. Rethlingshafer, New York: McGraw-Hill, pp. 510–546.

Robert, C. P. (1995), "Simulation of Truncated Normal Variables," *Statistics and Computing*, 5, 121–125.

Rosati, A. G., Stevens, J. R., Hare, B., and Hauser, M. D. (2007), "The Evolutionary Origins of Human Patience: Temporal Preferences in Chimpanzees, Bonobos, and Human Adults," *Current Biology*, 17, 1663–1668.

Spearman, C. (1904), "'General Intelligence,' Objectively Determined and Measured," *The American Journal of Psychology*, 15, 201–292.

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.

Verdinelli, I. and Wasserman, L. (1995), "Computing Bayes Factors Using a Generalization of the Savage-Dickey Density Ratio," *Journal of the American Statistical Association*, 90, 614–618.

Vlamings, P. H., Hare, B., and Call, J. (2010), "Reaching Around Barriers: The Performance of Great Apes and 3–5 Year-old Children," *Animal Cognition*, 13, 273–285.

Yuan, Y. and Johnson, V. E. (2011), "Goodness-of-fit Diagnostics for Bayesian Hierarchical Models," Working Paper 69, UT MD Anderson Cancer Center Department of Biostatistics Working Paper Series, available at *http://www.bepress.com/mdandersonbiostat/paper69* (accessed June 2011).

APPENDIX A

APPENDIX

Posterior sampling using MCMC depends on the chain converging to the target (posterior) distribution and on adequate mixing of the chain. The selected model for primate intelligence has species effects, species*paradigm effects, and assumes that the variance of the error terms is constant across assessments. Inference from the selected model uses a longer run of the MCMC algorithm than was used for preliminary analyses to choose a model because more precision is required for posterior inferences and comparisons across species. Two lengthy runs were used to check that each arrived at the same distribution when initialized at different points of the parameter space. The latter run was used for inference, and consisted of a burn-in period of 250,000 iterations and an additional run of 7,500,000 iterations after burn-in.

Trace plots are now included for select model quantities using the longest run. Only every 1000^{th} iteration is shown, starting from initialized values. The unthinned sequences often exhibit extremely strong autocorrelation and persistent dependence and consequently we used a long MCMC run. However, as demonstrated in the plots, the thinned chain appears to behave as desired, with the trace plots suggesting stationary behavior with little autocorrelation remaining.

Tuning parameters were allowed to change through the first 200,000 runs of a total burn-in period of 250,000 runs. After initialization of the tuning parameters, acceptance rates from each consecutive block of 300 runs were calculated, and moderate adjustments were made to increase or decrease the tuning parameters, as appropriate, if the block had too many or too few acceptances. For updates of individual parameters, the target rate was 35–45%, and for the vector of variance parameters

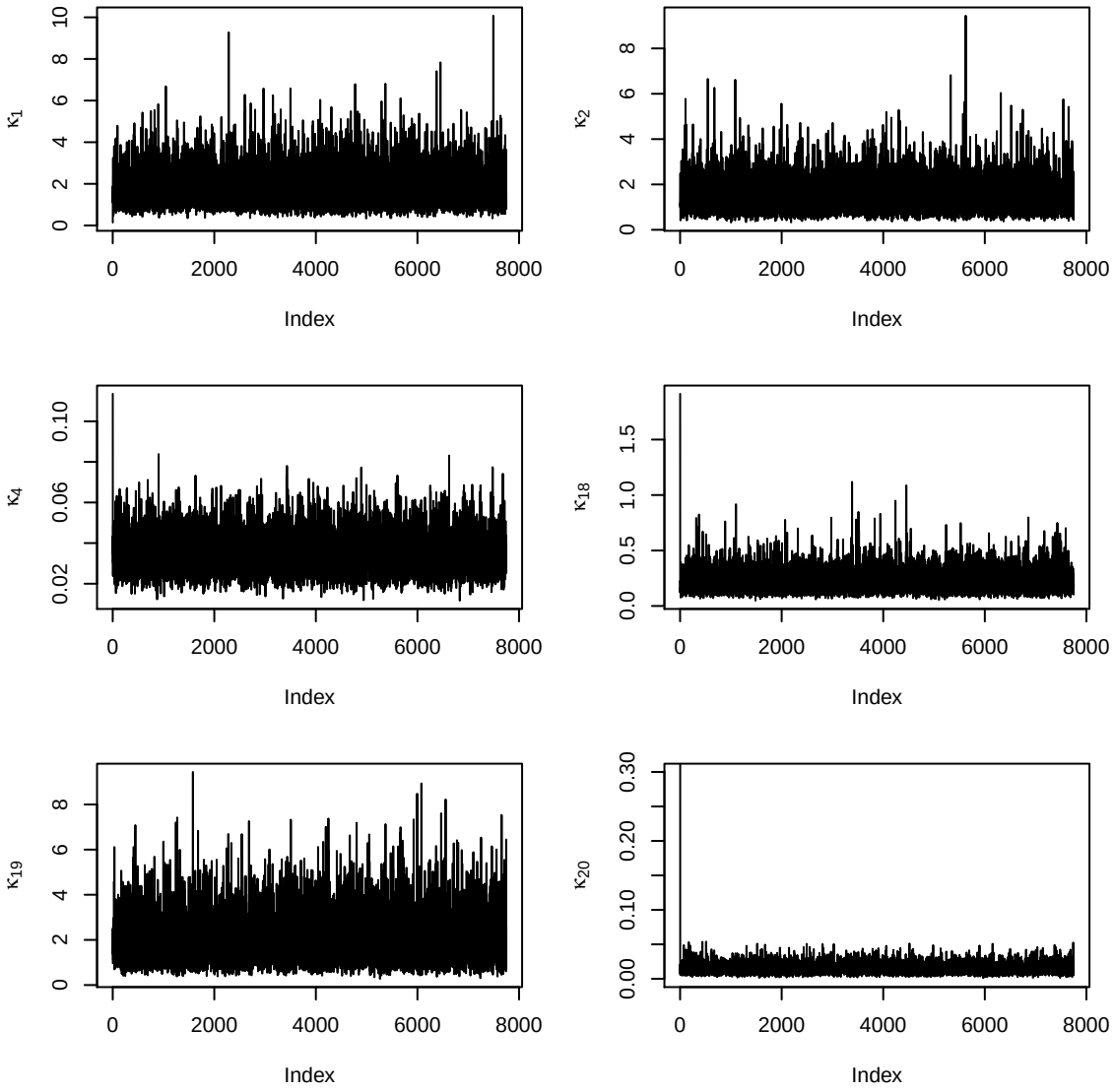

Fig. 10. Trace plots for each κ_j from a rank response assessment. Each depicted sequence has been thinned to include every 1000^{th} iteration of the original sequence. The final 7500 iterations represent those after the burn-in period.

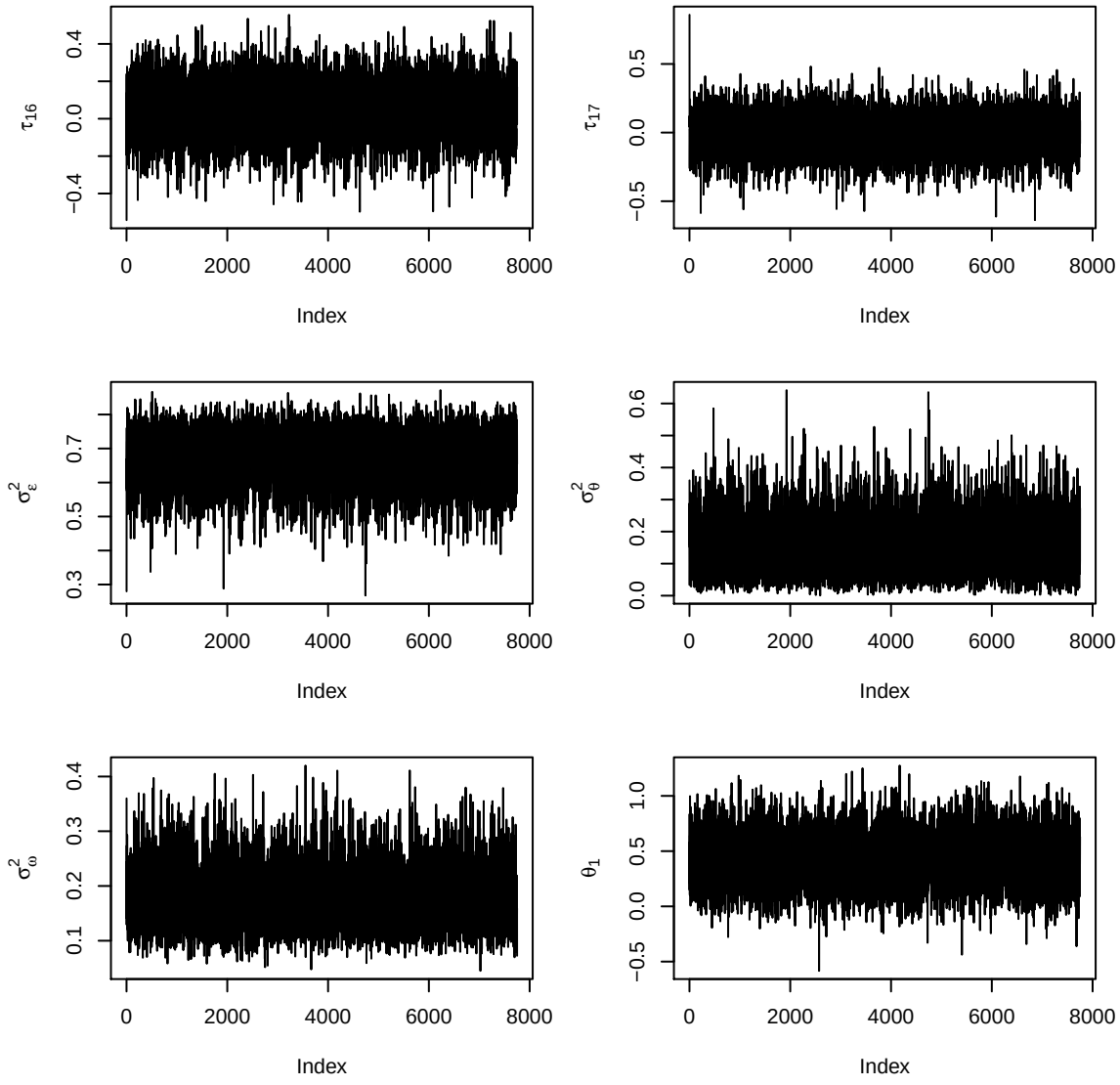

Fig. 11. Trace plots for: τ_j from the last two binomial response assessments; each of the three variance parameters; and the species effect for chimpanzees (species 1). Each depicted sequence has been thinned to include every 1000^{th} iteration of the original sequence. The final 7500 iterations represent those after the burn-in period.

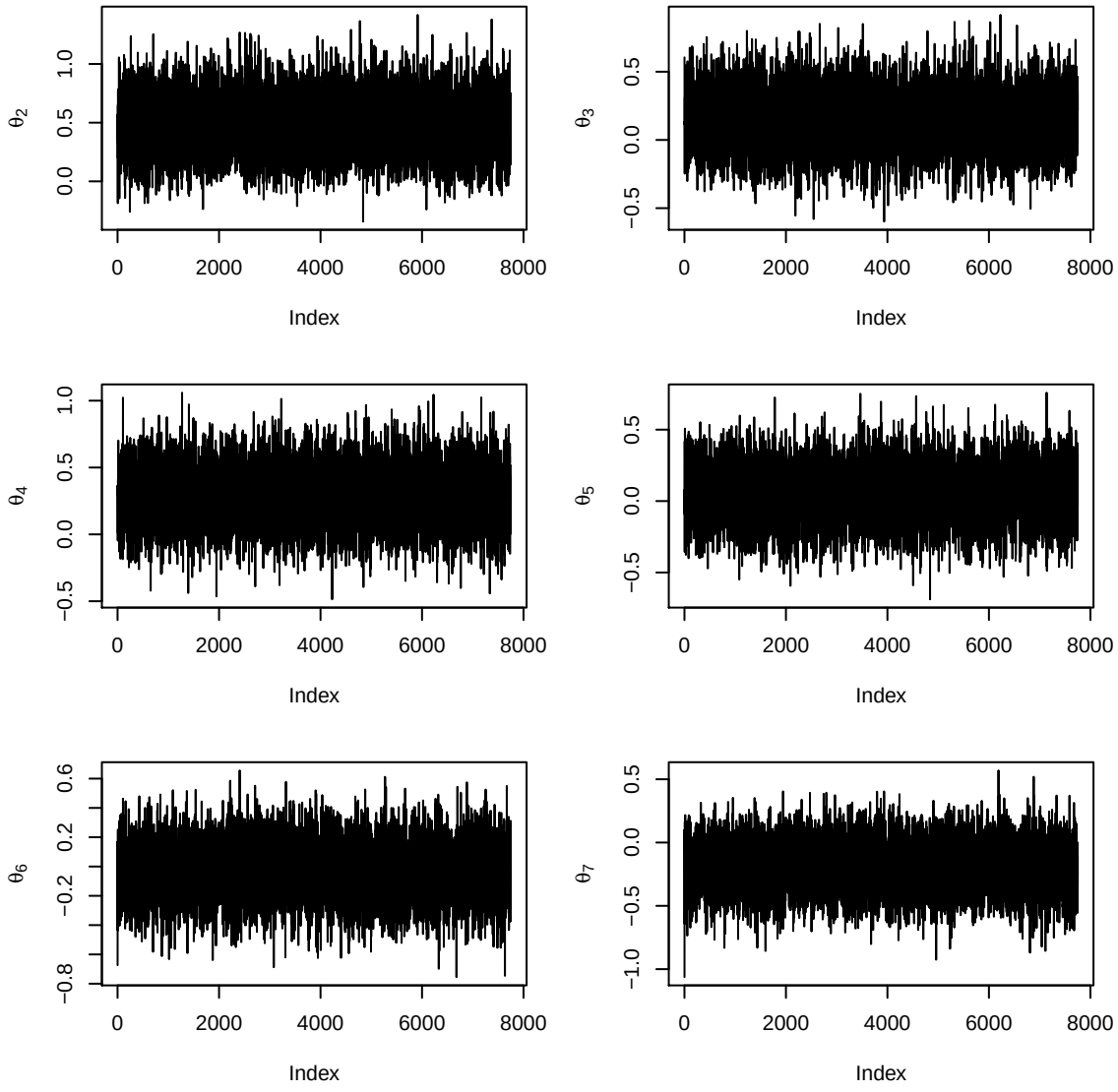

Fig. 12. Trace plots for the species effects for bonobos, gorillas, orangutans, spider monkeys, capuchin monkeys, and long-tailed macaques (species 2–7, respectively). Each depicted sequence has been thinned to include every 1000^{th} iteration of the original sequence. The final 7500 iterations represent those after the burn-in period.

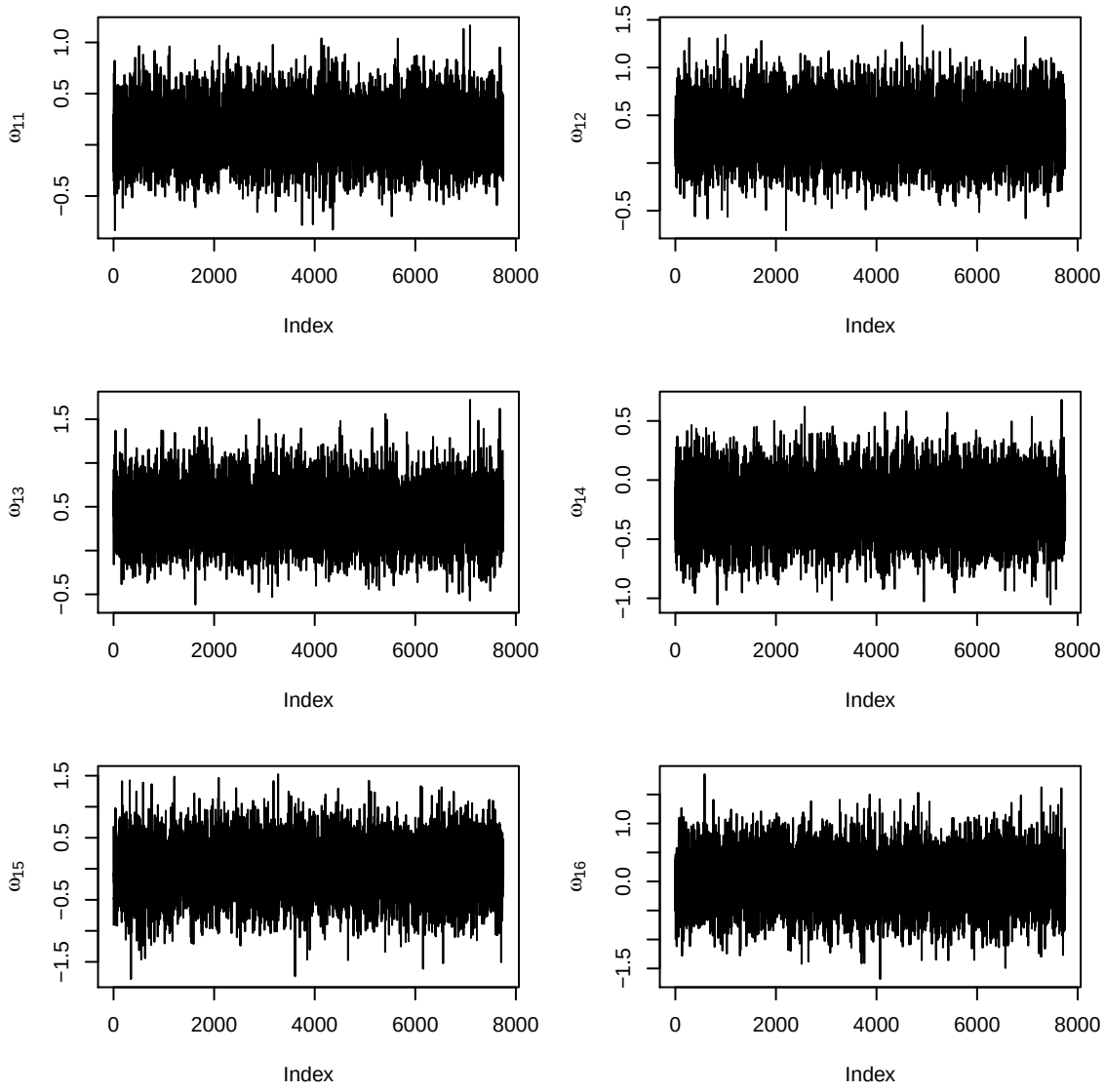

Fig. 13. Trace plots for the species*paradigm effects for chimpanzees (species 1) in paradigms 1–6. Each depicted sequence has been thinned to include every 1000^{th} iteration of the original sequence. The final 7500 iterations represent those after the burn-in period.

$\boldsymbol{\sigma}$ the target rate was 22-28%. Table XII shows the initial and final values for the tuning parameters. Two tuning parameters, b_{MH} and v^2, were given initial values but not allowed to change because inference should not be particularly sensitive to reasonable choices for their values.

Table XII. Values used for tuning parameters in the MCMC algorithm for ultimate inference on the selected model. After 200,000 iterations, the tuning parameters were no longer allowed to change. Target acceptance rates for each τ_j and each κ_j were 35–45%, while the target acceptance rate for $\boldsymbol{\sigma}$ was 22–28%. Two tuning parameters, v^2 and b_{MH}, were not permitted to change during the algorithm run.

Tuning Parameter (Quantity Affected)	Initial Value	Final Value	Tuning Parameter (Quantity Affected)	Initial Value	Final Value
a_{MH} $(\boldsymbol{\sigma})$	300.	203.	v_{10}^2 (τ_{10})	0.64	0.16
b_{MH} $(\boldsymbol{\sigma})$	0.05	0.05	v_{11}^2 (τ_{11})	0.64	0.27
v^2 $(\mathbf{z})$	0.09	0.09	v_{12}^2 (τ_{12})	0.64	0.02
c_1^2 (κ_1)	0.16	1.59	v_{13}^2 (τ_{13})	0.64	0.02
c_2^2 (κ_2)	0.16	2.11	v_{14}^2 (τ_{14})	0.64	0.02
v_3^2 (τ_3)	0.64	0.06	v_{15}^2 (τ_{15})	0.64	0.02
c_4^2 (κ_4)	0.16	0.29	v_{16}^2 (τ_{16})	0.64	0.02
v_5^2 (τ_5)	0.64	0.05	v_{17}^2 (τ_{17})	0.64	0.02
v_6^2 (τ_6)	0.64	0.05	c_{18}^2 (κ_{18})	0.16	0.98
v_7^2 (τ_7)	0.64	0.05	c_{19}^2 (κ_{19})	0.16	2.11
v_8^2 (τ_8)	0.64	0.09	c_{20}^2 (κ_{20})	0.16	0.83
v_9^2 (τ_9)	0.64	0.07			

VITA

Bradley John Barney attended Eastern Arizona College and Brigham Young University as an undergraduate student. He graduated twice from Brigham Young University, first with a Bachelor of Arts in economics (August 2003) and then with a Master of Science in statistics under the direction of Dr. H. Dennis Tolley (December 2007). He graduated with a Doctor of Philosophy in statistics from Texas A&M University under the direction of Dr. Valen E. Johnson and Dr. Simon J. Sheather in August 2011. His email address is bjbstat@gmail.com (preferred communication medium) and his mailing address is: Bradley Barney, c/o Dr. Simon J. Sheather, Department of Statistics, TAMU, MS 3143, College Station, TX 77843.

The typist for this dissertation was Bradley Barney.