MISSING IN ACTION: *

A BAYESIAN HIERARCHICAL MODEL FOR NA/DK RESPONSES IN SURVEYS

Adam Ramey†

August 6, 2009

Abstract

The problem of missing data is virtually endemic to political science research. Popular imputation techniques (e.g. Rubin 1987, King et al. 2001) have become widespread and are incredibly useful. However, there are many instances when modeling the mechanism of missingness directly will improve over both of these extant techniques in terms of model fit and, more importantly, theoretical microfoundations. This paper seeks to fill this gap in the methodological literature, with a particular emphasis on survey data. Specifically, I propose a new approach to modeling NA/DK responses in ordinal survey questions as the products of choice on the part of respondents. Drawing insights from the marketing literature (Bradlow and Zaslavsky 1999), I present a Bayesian hierarchical model that treats responses as the product of multiple latent variables: saliency, opinion, and decisiveness. Estimation is performed by MCMC methods, employing the data augmentation technique of Tanner and Wong (1987) due to the lack of closed-form conditional posterior distributions. An application to citizens' perception of candidate ideology is presented. The results provide both new and different insights that are missed by extant methods (e.g., listwise deletion and multiple imputation).

Keywords: Hierarchical Models, Missing Data, MCMC, Data Augmentation, Survey Research

*I would like to thank Curt Signorino, Michael Peress, Kevin Clarke, Jeremy Kedziora, Shawn Ramirez, Yoji Sekiya, and Jun Xiang for comments. I would also like to thank my wife Joslyn Ramey for suggestions and comments on the application of the method. All remaining errors are my own.

†Department of Political Science, University of Rochester, Rochester, NY 14620. E-mail: adam.ramey@rochester.edu

1 Introduction

Political science stands in a unique position amongst the other sciences. On the one hand, our discipline lacks the ability to exercise true experimental control. Physicists and chemists are able to manipulate parameters in their models with relative ease, making ensuing statistical analyses both theoretically and practically more tractable. Political science, with its unit of analysis typically set to either individuals or nations, attempting to impose control is either impossible, immoral, or both.

On the other hand, unlike natural scientists, political scientists are able to *ask* our units of analysis about their motivations and beliefs via surveys. Indeed, the aid of surveys is indispensable when it comes to probing public opinion. Surveys are relatively inexpensive (vis-à-vis quantum mechanical experiments) and afford the researcher a tremendous opportunity to gain insight into the dynamics of politics and decision-making.

Despite their flexibility and usefulness, surveys often present more problems than they solve. Low response rates have the possibility of biasing any sort of statistical inference that is attempted, though recent scholars (e.g., Peress 2007) have proposed remedies. Perhaps more difficult to address is the presence of missing data among those who do respond to the survey. Many surveys (e.g., the American National Election Study, General Social Survey, the Eurobarometer) give individuals the opportunity to skip questions (NA) or elicit "don't know (DK)" responses. Though traditional practice usually involved deleting these observations listwise or pairwise, it seems rather clear that doing so can lead to biased inferences down the road. As such, methods are required to address this problem.

Bradlow and Zaslavsky (1999) present just such an approach—one that models the missingness by using a hierarchical, multiple latent variable approach.

This method considers individuals' responses to ordinal indicators as a product of saliency, strength of opinion, and decisiveness (when opinion is weak). In this paper, I present the framework that was developed by Bradlow and Zaslavsky (1999) and demonstrate how it may be applied to the prevalent problem of NA's in American politics survey research. In Section 2, I discuss the existing literature on missingness in general and multiple imputation in particular. Section 3 presents the essential features of model as derived by Bradlow and Zaslavsky. Section 4 discusses the estimation strategy. The chosen technique is the data augmentation technique of Tanner and Wong (1987; see also Jackman 2000 and Albert and Chib 1993). Last, Section 5 applies the methodology of this paper to study respondents' perception of Congressional candidates' ideology in the 1992 general election.[1]

2 Literature

Missing data has long drawn the ire of researchers throughout the social sciences. In the case of surveys, known for their low response rates to begin with, item non-response can cause serious problems for multivariate analysis. Moreover, traditional remedies like listwise deletion or mean-insertion have been shown to cause serious bias in estimates and/or inferences (see, e.g., King et al. 2001). As a result, a large literature in both applied statistics (Rubin 1976, 1977, 1987; Gelman et al. 2004; Gelman, King, and Lin 1999) and political science/econometrics (King et al. 2001; Berinsky 1999; Brehm 1997; Heckman 1976) has emerged to model misingness in ways that minimize the inference bias that traditional remedies might induce.

This broad literature can be divided into two loosely-defined classes. The first

[1]For scholars interested in Monte Carlo-like analysis, Appendix A provides Monte Carlo evidence demonstrating that inferences from using an ordered probit model diverge widely from the hierarchical model presented in Section 3.

is the class of multiple imputation models (Rubin 1976, 1977, 1987; Gelman et al. 2004; Gelman, King, and Lin 1999). This paradigm seeks to "impute" the missing values by using other observed information in the data matrix. The method has become extremely popular, as it is relatively easy to implemement, as it is now commonly available in standard statistical packages like `STATA` and `R`. Some difficulty tends to arise when variables are nominal or ordinal, as is the case in surveys, but transformations can be applied to get around this. Nonetheless, imputation is about as close as one can get to a one-size-fits-all missing value methodology.

While this approach has wide applicability, it certainly has limitations. Two issues in particular may limit the usage and applicability of traditional imputation methods. First, the missingness must obey the so-called *missing at random* assumption. Following the notation of King et al. (2001), let D_{obs} denote observed data, D_{mis} denote missing data, D denote the total data, and M represent missingness. We say that data are missing-at-random (MAR) if $P(M|D_{obs}, D_{mis}) = P(M|D_{obs})$. More simply, data satisfy MAR if the missingness can be modeled as a function of observed data. A canonical example of this sort of missingness is the case of high wage-earners who are reluctant to report their income in surveys. While the income may be unobserved, there are several known correlates (e.g., education) that are *not* missing. By conditioning on these observed values, we may model missingness straightforwardly using existing algorithms. However, if the missing observations do not have observed covariates that can predict them in the data, MAR is not satisfied and multiple imputation is no longer an option.

The second issue, one that is less statistical and more conceptual, is the nature of the missing values in the first place. This issue can be posited as follows: when considering NA/DK responses in surveys, particularly those on opinion-oriented questions, is it necessarily the case that missing values are simply censorings or

"accidents"? Perhaps it is the case that individuals who do not choose a response or elicit "don't know" are actually making a choice, a choice in the same sense as the other categories given to them. Indeed, if this is true, King and his co-authors concede that cases "...when 'no opinion' means that the respondent really has no opinion rather than prefers not to share information with the interviewer, should be treated seriously and modeled directly..." (King et al. 2001, 59).

The other class of models for missing data are deemed by King et al. (2001) as "application specific methods" (e.g., Brehm 1997; Bartels 1998; Berinsky 1999; Heckman 1976). These methods are diverse and, as such, do not follow neatly into one category. Generally speaking, these methods require a specific modeling of the missingness mechanism, which can be difficult and will certainly vary across applications. For example, two such approaches (Berinsky 1999; Heckman 1976) consider data in terms of the so-called selection model. In particular, Berinsky (1999) models citizens' NA/DK response on racial atttiude questions as a selection process, whereby those choosing NA/DK select themselves out of the sample. This approach is novel and yields to substantively interesting results, but it too presents limitations. First, this class of models require an exclusion restriction to ensure proper identification. Second, the model restricts missingness to result from only one initial selection. It is also plausible to think that missingness could be due to indifference; that is, citizens are indifferent between certain categories and this leads them to report NA/DK. The approach of Bradlow and Zaslavsky (1999) considered in this paper and presented in the next section provides scholars with an ability to model without either the deficiencies of the selection-based approach or the imputations approach.

3 The model

Figure 1 about here.

3.1 Latent variables

Let $i = 1, 2, ..., N$ denote the set of respondents to the survey.[2] For each individual i, y_i is his ordinal response to the item. Typically, these sorts of items involve five- or seven-point scales. If i skipped the item, it is assumed that his response is coded as either NA or DK. In a usual analysis of this data, ordered probit or logit is the most common technique employed. The probabilities of the various y_i's are modeled as a function of covariates X and cutpoints c_q. Estimation is either achieved by maximizing a likelihood function or, with assignment of priors, a sampling from posterior distributions.

The modeling approach employed herein can be seen as a generalization of the ordered probit. The main departure is the multiplicity of latent variables. In the ordered probit, the y_i's are viewed as manifestations of an underlying y_i^*, where the various ordinal values are determined by cutpoints on the underlying latent scale. The hierarchical approach of this paper views an individual's response as a product of three underlying latent processes: *saliency*, *opinion*, and *decisiveness*. *Saliency*, given by ψ_i, is the first latent factor in the decision-maker's process. If the item is not salient, $\psi_i < 0$ and the respondent will elicit a NA/DK response.

If the item is in fact salient, the next stage is the existence of a true *opinion*, ϑ_i. Respondents whose latent opinion is more extreme are assumed to have "true" or definitive opinions. The extremity of opinion here is defined in terms of cutpoints, c_q, where $q = 1, ..., Q-1$ represents the ordinal response category, $c_0 = -\infty$ and $c_Q = \infty$. An opinion ϑ_i is considered extreme if

$$\vartheta_i < c_L \tag{1}$$

[2]The model presented herein was developed in Bradlow and Zaslavsky (1999). It was originally designed to analyze multiple ordinal indicators. However, in this paper, the decision structure has been modified so as to accommodate analysis of only one ordinal indicator. Readers interested in the analysis of multiple issues may consult the original paper.

or

$$\vartheta_i > c_H, \tag{2}$$

where the cutpoints c_L and c_H depend on the number of possible ordinal responses given on the particular item. In particular, it is assumed that $c_L = c_{q_L - 1}$ and $c_H = c_{q_H}$. q_H and q_L are typically chosen so that they straddle the cutpoint that corresponds to the center-most category. For example, if the observed data is from a seven-point scale, category 4 is at the center. This makes $q_L = 3$ and $q_H = 5$ ideal candidates for for $c_L = c_2$ and $c_H = c_5$ respectively.[3]

Should i have a ϑ_i that satisfies either (1) or (2) above, we assume that he will elicit an ordinal response and not NA/DK. However, if $\vartheta_i \in [c_L, c_H]$, we say that i is in the *region of indifference*. This, in turn, leads to the last stage in the decision tree. We can imagine that a latent opinion in this range might lead to one of two observed behaviors. If the individual is decisive, then he would be more inclined to elicit an ordinal response than if he were indecisive. This notion is formalized in the third latent variable δ_i, where $\delta_i \geq 0$ implies i's decisiveness on the item and hence, an ordinal response will be given. If he is not decisive, $\delta_i < 0$ and the NA/DK response is given. The entire process is depicted in Figure 1.

A few comments are warranted before proceeding. First, this model provides a rich description of respondents' behavior. For example, the NA/DK response can be observed if the respondent is indifferent or if the item is simply not salient. These are very different kinds of NA's and are necessarily modeled as such. Second, if there are no NA's, this more complicated model reduces to a simple ordered probit. This model is consequently always the preferred option, as it can pick up effects that the ordered probit would miss, but still turn out the same results where NA's are nonexistent.

[3]The boundaries of this zone are not by any means set in stone. Depending on the data, researchers may desire to modify the indifference region bounds accordingly.

3.2 Hierarchy I: Distribution of $y|\phi_1, \phi_2$

It is assumed that the three latent variables, ψ_i, ϑ_i, and δ_i distributed normally with variance 1. Saliency, ψ_i, is assumed have a mean $\mu_i^{\psi_i}$ such that

$$\mu_i^{\psi_i} = \eta_i + X_i^{\psi}\beta^{\psi}, \tag{3}$$

where η_i is a random intercept that allows individuals to vary in terms of saliency and X_i^{ψ} is a vector of covariates thought to influence saliency. For the latent opinion ϑ_i, the mean is given by

$$\mu_i^{\vartheta_i} = X_i^{\vartheta}\beta^{\vartheta}, \tag{4}$$

where X^{ϑ} is a set of covariates thought to influence the latent opinion. The final latent variable, decisiveness δ_i has a mean

$$\mu_i^{\delta_i} = X_i^{\delta}\beta^{\delta}, \tag{5}$$

and X_i^{δ} is a vector of covariates affecting decisiveness.

These latent draws may thus be summarized as follows:

$$\psi_i \sim \mathcal{N}(\mu_i^{\psi_i}, 1) \tag{6}$$

$$\vartheta_i \sim \mathcal{N}(\mu_i^{\vartheta_i}, 1) \tag{7}$$

$$\delta_i \sim \mathcal{N}(\mu_i^{\delta_i}, 1). \tag{8}$$

Figure 2 about here.

The complete data likelihood is found based on the definitions of the various

latencies above. In particular, the likelihood function is given by

$$\mathcal{L}(\phi_1, \phi_2 | y_{ij}, X) \propto \prod_{i=1}^{N} p_i(\phi_1, \phi_2 | y_{ij}, X), \tag{9}$$

where the p_i are probabilities associated with the ordinal outcomes. To illustrate more simply how these are derived, Figure 2 presents a simplified decision tree that is broken down into three different probabilities: r, s, and t. To evaluate p_i in these terms, we need to look at the various responses that could be provided on the ordinal items. First, we consider the case where i elicits NA/DK. This could have resulted two ways, as is seen in Figure 2. Thus, the probability of observing a NA/DK response is

$$\begin{aligned} Pr(y_i = NA/DK) &= r + s(1-r)(1-t) \\ &= \Phi(-\mu_i^{\psi_i}) + (\Phi(c_{q_H} - \mu_i^{\vartheta_i}) - \Phi(c_{q_L} - \mu_i^{\vartheta_i}))(1 - \Phi(-\mu_i^{\psi_i})) \\ &\times \Phi(-\mu_i^{\delta_i}). \end{aligned} \tag{10}$$

Second, we look at non-NA responses that fall outside of the indifference region. The probability of observing a response outside of this region is given by

$$\begin{aligned} Pr(y_i = q \notin [q_L, q_H]) &= (1-r)(1-s) \\ &= (1 - \Phi(-\mu_i^{\psi_i}))(\Phi(c_q - \mu_i^{\vartheta_i}) - \Phi(c_{q-1} - \mu_i^{\vartheta_i})). \end{aligned} \tag{11}$$

Finally, there is the probability of observing a non-NA response that is within the indifference region. Examining Figure 2, this is given by

$$\begin{aligned} Pr(y_i = q \in [q_L, q_H]) &= (1-r)st \\ &= (1 - \Phi(-\mu_i^{\psi_i}))(\Phi(c_q - \mu_i^{\vartheta_i}) - \Phi(c_{q-1} - \mu_i^{\vartheta_i})) \\ &\times (1 - \Phi(-\mu_i^{\delta_i})). \end{aligned} \tag{12}$$

We can assemble all of these into a single statement as follows:

$$p_i(y_i) = \begin{cases} \Phi(-\mu_i^{\psi_i}) + (\Phi(c_{q_H} - \mu_i^{\vartheta_i}) - \Phi(c_{q_L} - \mu_i^{\vartheta_i}))(1 - \Phi(-\mu_i^{\psi_i})) & \\ \qquad \times \Phi(-\mu_i^{\delta_i}), & \text{if } y_i = NA/DK \\ (1 - \Phi(-\mu_i^{\psi_i}))(\Phi(c_q - \mu_i^{\vartheta_i}) - \Phi(c_{q-1} - \mu_i^{\vartheta_i})), & \text{if } y_i = q \notin [q_L, q_H] \\ (1 - \Phi(-\mu_i^{\psi_i}))(\Phi(c_q - \mu_i^{\vartheta_i}) - \Phi(c_{q-1} - \mu_i^{\vartheta_i})) & \\ \qquad \times (1 - \Phi(-\mu_i^{\delta_i})), & \text{if } y_i = q \in [q_L, q_H] \end{cases} .$$

3.3 Hierarchy II: Distribution of $\phi_1|\phi_2$

The second layer of hierarchy in this Bayesian model looks at the vector of prior parameters $\phi_1 = \eta$. Each of these parameters is normally distributed as follows:

$$\eta_i \sim \mathcal{N}(X_i^\eta \beta^\eta, \sigma_\eta^2). \tag{13}$$

X^η is a matrix of covariates that is assumed to influence the saliency of individuals in the data set. In the application presented subsequently, I assume that the matrices of covariates for ϑ, δ, and η are the same.

3.4 Hierarchy III: Distribution of ϕ_2

The third and final layer of hierarchy in this model is the vector of hyperparameters for the coefficients, variances, and cutpoints: $\phi_2 = (\beta^\psi, \beta^\delta, \beta^\vartheta, \beta^\eta, \sigma_\eta^2, \boldsymbol{c})$. Each group of regression coefficients for latent parameter Z are drawn from a Multivariate Normal distribution with mean 0 and covariance V:

$$\beta \sim \mathcal{MVN}(0, V). \tag{14}$$

The choice of V can be as large or as small as the researcher desires. For the variance of the random saliency intercept, I employ an uninformative conjugate prior

that is defined on the positive reals. Standard results show such a distribution to be the Inverse Gamma. Thus, the prior for σ_η^2,

$$\sigma_\eta^2 \sim \mathcal{IG}\left(\frac{\rho}{2}, \frac{1}{2}\right), \tag{15}$$

and ρ is chosen to be 0.5.

Last, and perhaps the trickiest, is the vector of cutpoints. All cutpoints are assumed to be drawn uniformly from the last cutpoint to the current cutpoint. More specifically,

$$c_q | c_{q-1}, c_{q+1} \sim U(c_{q-1}, c_{q+1}), q = 1, 2, ..., Q-1, \tag{16}$$

$c_0 = -\infty$, $c_Q = \infty$, and, for identification, some $q' \in \{1, 2, ..., Q-1\}$, $c_{q'} = 0$.

3.5 The full posterior

We can combine the expressions for the likelihood and priors above to form the complete posterior:

$$\pi(\phi_1, \phi_2 | y, X) \propto \mathcal{L}(y | \phi_1, \phi_2, X) p(\phi_1 | \phi_2, y, X) p(\phi_2). \tag{17}$$

As is the case with all hierarchical models, there is a very large number of parameters to estimate. Indeed, we are required to estimate a minimum of $N + Q - 1$ parameters (not including the β's or σ^2). The choice of covariates above this number makes things even more complicated. With the number of parameter estimates far exceeding the sample size, we must consider alternate ways to sample from equation (17).

4 Estimation

4.1 Problems with Maximum Likelihood and standard Gibbs sampling

Prior to proceeding with a description of how to estimate this model using Bayesian techniques, it is important to say a word on why Maximum Likelihood Estimation (MLE) is not a feasible alternative. Some scholars who are uncomfortable with the notion of Bayesian priors (especially those chosen based on conjugacy and not on knowledge) would be more satisfied with using Frequentist analogs of Bayesian models. In the case of the complex hierarchical model presented heretofore, it would simply not be practical or even possible to estimate the model via MLE.

To understand the situation better, consider the familiar example of ideal point estimation. The two dominant approaches are Pool and Rosenthal's (1997) MLE-based DW-NOMINATE and Clinton, Jackman, and Rivers' (2004) Bayesian IRT model. If M is the number of roll calls and L is the number of legislators, there are at least $M + L$ parameters to estimate (two cutpoints per roll call and one ideal point per legislator). To make the estimation feasible, Poole and Rosenthal must estimate parameters in parts, fixing the M cutpoints while estimating the L ideal points and then vice versa, iterating until convergence. In contrast, the Bayesian approach is computationally simple and can be written and run in a few moments in `WinBUGS` or `MCMCpack` (Martin and Quinn 2007).

In the model presented in this paper, there are even *more* parameters to estimate than in the ideal point model. Indeed, the draws of the latent ϑ alone are essentially a one-dimensional ideal point model. With the addition of the other layers of hierarchy, the Bayesian approach is simply much more tractable.

This is not to say, however, that standard Gibbs sampling is without its difficulties. The Gibbs sampler requires the derivation of the conditional distribu-

tions, many of which do not exist in closed form due to the inter-connectedness of various layers of the hierarchy and the complexity of the likelihood. Metropolis steps could be used in these cases, but the data augmentation strategy of Tanner and Wong (1987; see also Albert and Chib 1993 or Jackman 2000) allows straightforward Gibbs sampling.

4.2 Data-augmented Gibbs sampling and conditional posteriors

The data augmentation strategy involves sampling from latent variables ψ, ϑ, and δ. Conditional on these sampled parameters, the conditional distributions of the remaining parameters exist in closed form and, hence, a Gibbs sampling routine may be employed. Let us first present the conditional distributions for the three latent variables. All of these results follow clearly from the decision tree presented in Figure 1. The distribution of saliency ψ can be summarized as follows

$$\psi_i|\phi_1,\phi_2,X,y \sim \begin{cases} \mathcal{TN}_{\Re_+}(\mu_i^{\psi_i},1), & \text{if } y_i = q \\ \mathcal{N}(\mu_i^{\psi_i},1), & \text{if } y_i = NA/DK, \vartheta_i \in [c_L,c_H], \delta_i < 0 \\ \mathcal{TN}_{\Re_-}(\mu_i^{\psi_i},1), & \text{else} \end{cases},$$

where $\mathcal{TN}$ is the truncated normal distribution and the subscript $\Re_+$ restricts the distribution to the positive reals. The first case states that non-NA/DK responses are drawn with all positive values, as the item is salient. The second case allows NA/DK responses from the indifference region to be either positive or negative. The final case restricts all other NA/DK to be negative, as the question was definitely not salient.

The conditional distribution of the latent opinion ϑ_i follows in a similar fashion:

$$\vartheta_i|\phi_1,\phi_2,X,y \sim \begin{cases} \mathcal{TN}_{(c_{q-1},c_q)}(\mu_i^{\vartheta_i},1), & \text{if } y_i = q \\ \mathcal{TN}_{(c_L,c_H)}(\mu_i^{\vartheta_i},1), & \text{if } \psi_i > 0, y_i = NA/DK \\ \mathcal{N}(\mu_i^{\vartheta_i},1), & \text{else} \end{cases},$$

where the the first two conditions are drawn from Normal distributions truncated at relevant cutpoints. Non-NA/DK responses are truncated to lie between their previous and current cutpoint. Salient questions that yield NA/DK's must come from the region of indifference, $[c_L, c_H]$. All other NA/DK responses are untruncated.

The conditional distribution of decisiveness, δ_i is given by

$$\delta_i|\phi_1,\phi_2,X,y \sim \begin{cases} \mathcal{TN}_{\Re_+}(\mu_i^{\delta_i},1), & \text{if } y_i = q, \vartheta_i \in [c_L,c_H] \\ \mathcal{TN}_{\Re_-}(\mu_i^{\delta_i},1), & \text{if } y_i = NA/DK, \vartheta_i \in [c_L,c_H], \psi_i \geq 0 \\ \mathcal{N}(\mu_i^{\delta_i},1). & \text{else} \end{cases}.$$

In the first case, an ordinal response is given but the latent opinion is in the region of indifference, so the individual must have been decisive. The second case is also in the region of indifference, but the person elicits a NA/DK, so he was not decisive. For all other case, the normal distribution is untruncated.

Once we condition on the latent variables described above, the distributions of the others follow easily. This comes as a results of the Normal and Inverse Gamma priors chosen previously for the data generating process and the priors. To make notation simpler, let $\Omega = (\psi, \vartheta, \delta)$. We first consider the parameter at the first level of hierarchy, ϕ_1. The conditional distribution is given by

$$\eta_i|\Omega,\phi_2,X,y \sim \mathcal{N}\left(\frac{1}{1+\frac{1}{\sigma_{\eta^2}}}\left((\psi_i - X_i^{\psi}\beta^{\psi}) + \frac{1}{\sigma_\eta^2}X_i^{\eta_i}\beta^{\eta}\right), \frac{1}{1+\frac{1}{\sigma_{\eta^2}}}\right) \tag{18}$$

The parameters in ϕ_2 are even easier to derive, once again a function of con-

jugacy. In fact, all β's and the σ^2 have conditional posteriors of a similar form. The β's are each Multivariate Normally distributed, combining the prior additively with the regression-based estimate. Since they all have the same form, I will present the conditional distribution only once for a generic parameter m. If Z is the relevant response variable, the conditional distribution of β^m is given by[4]

$$\beta^m|\Omega,\phi_1,\phi_2 \sim \mathcal{MVN}\left(\left(\frac{1}{\sigma_m^2}X_m^TX_m+\frac{1}{V}I\right)^{-1}\frac{1}{\sigma_m^2}X_m^TZ,\left(\frac{1}{\sigma_n^2}X_m^TX_m+\frac{1}{V}I\right)^{-1}\right) \quad (19)$$

and the distribution for σ_η^2 is

$$\sigma_\eta^2|\Omega,\phi_1,\phi_2 \sim \mathcal{IG}\left(\frac{N+\rho}{2},\frac{(\eta-X^\eta\beta^\eta)^T(\eta-X^\eta\beta^\eta)+1}{2}\right). \quad (20)$$

The final conditional distribution to consider is that of the cutpoints, c_q. As has been shown in Albert and Chib (1993), the conditional distribution of the cutpoints will be uniform, but there must be a correction to ensure proper and sensible estimates. In particular, for each c_q,

$$c_q|c_{q-1},c_{q+1} \sim U(\max\{c_{q-1},\max_{y_i=q}\vartheta_i\},\min\{c_{q+1},\min_{y_i=q+1}\vartheta_i\}). \quad (21)$$

If we examine (21) more closely, the max and min conditions are designed to ensure finiteness of the draws of cutpoints. For example, to draw a cutpoint c_1, we would need c_0 and c_2. Since $c_0 = -\infty$, we need to obtain an endpoint that is finite. Hence, we take the maximum of $-\infty$ and a value of the latent opinion, which by definition will always be the choice. A similar line of argument applies to c_{Q-1}. Last, for identification, I fix $c_1 = 0$.

[4]For all parameters except η, $\sigma_m^2 = 1$ based on the definitions above.

5 Application: Perceptions of Candidate Ideology

5.1 Problem and Data

The literature on public opinion and voting behavior as far back as Converse (1964) has been concerned with citizens' ability to make ideological judgments. From this motivation, a large literature has emerged examining citizens' perception of candidates' ideology in various contexts (Feldman and Conover 1983; Conover and Feldman 1989; Delli Carpini and Keeter 1993; Rahn 1993; Sigelman et al. 1995; McDermott 1997; Althaus 1998; Koch 2000; Koch 2002). Indeed, the issue of citizens' abilities to place themselves and candidates in an ideological space is of great relevance to both scholars of voting behavior and those of formal voting theory.

To measure citizen placements, most scholars employ questions from large scale surveys like the American National Election Studies (ANES). In these surveys, respondents are given opportunities to place themselves and candidates on seven-points scales ranging from "extremely liberal" (1) to "extremely conservative" (7). Unforunately, citizens often are unable to place themselves or candidates, evidenced by high proportions of NA/DK responses. For example, in the 1992 ANES (Miller et al. 1993), approximately 31% of respondents are unable to place themselves on the seven-point scale. Even worse is the proportion of respondents able to place Congressional candidates. In that same ANES, approximately 56% of respondents could not place their Republican House candidate. While not surprising given that scholars have found that many Americans do not even know their incumbent Congressmen, let alone a candidate (Delli Carpini and Keeter 1997), this high proportion of NA/DK responses causes problems for any possible statistical inferences.

The hierarchical model presented above can be applied to the study of this data in a straightforward manner. For comparative purposes, models employ-

ing ordered probit with listwise deletion, as well as ordered probit with multiple imputation will be considered.[5] The dependent variable in all models will be the respondent's placement of the Republican House candidate in 1992 on the seven-point liberal-conservative scale. In the listwise deletion model, all NA/DK observations will be removed. For the imputation case, five imputed datasets are obtained using the `Amelia II` package in `R` (King et al. 2007). The hierarchical model treats the NA/DK cases as products of either lack of saliency or indeciveness, as defined in the model above.

The regressors considered in for the opinion matrix (X^{ϑ}) will be the same as those considered in the listwise deletion and the multiple imputation models. All of these variables are demographics that are considered important in the literature. The first group are basic demographics. *Age* is the age of the respondent in years. *Black* is a dummy indicating whether the respondent is black. *Male* indicates whether or not the respondent was a male. *Education* is the number of years in school for the respondent.

The other set of regressors are ideological in nature. *Republican* and *Democrat* are indicators for whether the respondents are self-identified Republicans or Democrats, respectively. *Extremism* is a measure of ideological extremism of the respondent. This is measured by transforming the self-reported 7-point scale placement as follows:

$$Extremism = |4 - Self|. \tag{22}$$

Since 4 is the mid-point of the scale, an individual choosing this category is given an extremism score of zero. However, individuals at either 1 or 7 will get an

[5]While tempting, the comparison with selection models is not considered here for a few reasons. First, varying exclusion restrictions can change the results of those models and, as a consequence, make comparability across models difficult to ascertain. Second, the dominant approaches in the literature are either listwise deletion or multiple imputation, making these two the appropriate baselines to compare the hierarchical model with.

extremism score of a 3. The motivation for rescaling ideology in this fashion is the idea that ideological extremists are probably more likely to place a Republican candidate to the right than moderates.[6]

In the hierarchical model, regressors for other levels of the model need to be specified. For simplicity, I employ the same regressors across all layers except saliency—i.e., $X^{\vartheta} = X^{\delta} = X^{\eta}$.[7] In the case of saliency, I simply consider the case of two predictors of saliency: level of political information and ideological extremism, as defined in equation (26).[8] Numerous scholars (e.g., Converse 1964; Zaller 1992) have enunciated the importance of political knowlede in citizens' abilities to make political judgments and use ideological scales. Measures of political knowledge are plentiful (Zaller 1992; Delli Carpini and Keeter 1993, 1997; Althaus 1998), and by far the most common is an additive index of correct answers to factual political questions.

To establish whether the additive index was indeed appropriate, I used the 1992 NES to obtain all components of the index used by Althaus (1998).[9] Rather than proceed immmediately, I performed a simple Principle Components Analysis on the data matrix, examining the factor loadings. All components of the index load in the same direction on the first Principal Component, which itself dominates all subsequent components in terms of variance explained. These results indicate that one the data can be largely explained by the first component. As such, I retain the factor scores for each individual on this first dimension and then correlate the factor scores with the simple additive index. The two correlate

[6]The more traditional way to do this is to put two terms into the regression: *ideology* and $ideology^2$. Thus, the polynomial term is able to capture the non-monotonic relationship discussed in this paragraph. This specification was attempted for this application and the results do not change. By opting for the simple linear extremism variable, interpretation of coefficients is more easily accomplished.

[7]The exception is X^{η}, which has two fewer regressors because they are included in X^{ψ}.

[8]I have also run the model allowing more regressors. Results do not change substantially. The choice of two regressors here is for simplicity of exposition.

[9]See Althaus (1998) for details. Components include ability to identify major political figures (e.g., the Vice President, Speaker of the House), functions associated with particular branches, ideological bent of major political parties, control of Senate and House by party, and the like.

at an extremely high 0.98. Since this is the case, I opt for the simpler additive index as a measure of knowledge, as incrememnts in these counts are more easily interpretatble than increments in factor scores.

To examine the role of political knowledge in terms of decisiveness and opinion, the variable $Info$ is included in these levels as well.

5.2 Estimation and Results

The hierarchical was estimated with the regressors as defined above over three chains with overdispersed starting values. Each chain was allowed to run for 100,000 iterations, discarding the first-half as burn-in and saving every 20th iteration due to storage limitations. To examine convergence, the Gelman and Rubin (1992) and Geweke (1992) diagnostics were employed. The chains were determined to have converged.

For the listwise deletion and multiple imputation cases, the `MCMCoprobit` function of Martin and Quinn's (2007) `MCMCpack` were employed for estimation. As described above, the imputed datasets were estimated using King et al.'s `Amelia` library for `R`.

Tables 1-3 about here.

Results from all three models are found in Tables 1-3. In each case, the posterior means and standard deviations are provided. When the 95% region of highest posterior density (HPD) lies on the same side of zero, coefficients are placed in bold. For the hierarchical model, results are further divided by level of hierarchy in Table 1. In the first equation, we see that both forms of party identification are positive predictors of latent opinion. This is not surprising, as self-identified partisans are more likely to pay attention to the campaign in the first place. Similarly, ideological extremism and level of political information also push the latent

distribution to the right. All other regressors have HPD's that cross zero, giving scholars less confidence that the effects are different from zero.

The other equations also provide valuable information. In the deciveness equation, we see that education has a strong effect on latent deciveness, demonstrating that increased education leads individuals in the region of indifference to have an increased probability of choosing an ordinal response. The saliency intercept (η) equation reveals that individuals' saliency intercepts are significantly influenced by respondent age. Last, the saliency equation reveals that increasing levels of information and ideological extremism have strong effects on the saliency of the placement question.

Given the information acquired from the hierarchical model in Table 1, we may now compare the results from the listwise-deleted and multiply imputed cases and examine how inferences change. For listwise deletion in Table 2, the changes are noticeable. First, the standard deviations for all regressors are larger than those in the hierarchical model. Second, and more importantly, *Age* and *Black* are found to have HPD's on the same side of zero in the listwise deletion model, indicating a "significant" impact on latent opinion. Further, while being a Republican has an effect on latent opinion, the effect of being a Democrat has an HPD that crosses zero. All three of these results contradict the results in Table 1. Age and race were not significant predictors, yet Democratic identification was. Thus, scholars employing listwise deletion would have come to incorrect conclusions regarding the effects of these three key demographics on individuals' opinions.

The results from multiple imputation in Table 3 are not much better from an inferential stand. *Age* and *Black* are revealed to be "siginificant" predictors of latent opinion, though the magnitude of effect for *Black* is much smaller. As was the case in the hierarchical model, *Democrat* is found to be a "significant" predictor, unlike the result in the listwise deletion case. Last, and still important, the

coefficient on $Info$, though still a significant predictor, is greatly reduced in magnitude. In turn, this would lead scholars to imply a less substantial role for levels of information in candidate placement.

In sum, the results from this application are clear. Employing listwise deletion or multiple imputation, due to the nature of the missing data, can lead to false positive inferences (i.e., *Age* and *Black*) and, in the case of listwise deletion, false negatives (i.e., no effect for *Democrat*). Moreover, the hierarchical model is able to go a step further and look at the factors that influence the saliency and deciveness factors across individuals. In turn, this helps scholars to develop a better understanding of the processes that lead to the NA/DK response in public opinion data.

6 Conclusion

This paper introduces the multiple latent variable approach to dealing with NA/DK responses in surveys pioneered by Bradlow and Zaslavsky (1999). The model was developed in detail and was shown to produce more accurate inferences than the ordered probit model. Though the method is computationally intensive and more difficult to implement than standard ordered probit or IRT models, the leverage gained is considerably worth the extra effort. By modeling the process that leads to NA/DK responses, we do not discard the observations and actually estimate parameters for them. Moreover, we do not simply impute the responses as if the NA/DK are missing at random. Indeed, this would have caused bias if the NA/DK responses are not random but instead deliberate.

The method can be extended to even more settings than the simple example presented herein. Scholars can use surveys with several ordinal indicators of the same sort (i.e., all 5- or 7-point scales) and examine cross-item effects as well as individual effects. The method works especially well when the covariates

are chosen based on strong theory. The hierarchical approach allows scholars to more realistically model the data-generating process and, hence, generate more reliable estimates of the effects. Further effort is needed to improve computational efficiency of this approach. Nonetheless, this approach is invaluable for survey researchers who seek to understand public opinion but have, until now, been restrained by an overabundance of NA/DK responses.

Appendix A: Data Experiment—When Ordered Probit Goes Wrong

To explore the features of this estimator, I construct a simple Monte Carlo experiment. In this experiment, some data will be generated according to the data-generating process described in the model and the $\hat{\beta}$'s will be estimated. The results from this estimation will be compared with an ordered probit that omits the missing values in a listwise fashion.[10]

Before proceeding, a few cautionary words are in order. First and foremost, the mere idea of a Monte Carlo experiment is a bit of an oddity in the Bayesian statistical framework. Monte Carlo analyses are typically used to analyze large-sample properties of estimators. In the Bayesian framework, large-sample properties are not usually of interest. Second, the way in which a Monte Carlo proceeds is in itself a violation of Bayesian philosophical principles. Monte Carlos usually begin with declaring knowledge of "true parameters," generating data thereafter with this knowledge. If the estimator at hand is unbiased, for a very large number of repetitions, the Monte Carlo should return a distribution whose mean is the "true" parameters set in the first place.

For a Frequentist, this setup is perfectly legitimate. The "true" parameter θ is assumed *a priori* to be some fixed value. One generates estimates $\hat{\theta}$ which vary due to randomness. For a Bayesian, θ does not exist as a single point that is out there in the world. Rather, θ (and not $\hat{\theta}$) is a distribution. This of course presents a philosophical conundrum for the Bayesian.

I approach these matters with a pragmatic mindset. On the one hand, I accept the philosophical barrier that Monte Carlo analysis presents to the Bayesian framework. On the other hand, however, it is extremely useful to demonstrate

[10]This experiment was also run for the case in which multiple imputation was employed in the ordered probit model. The distributions of coefficients, as well as predicted probabilities, very nearly match the results from the ordered probit model. As such, the plots are not distinguishable and are not presented. Results from this approach are available upon request.

the estimator's ability to recover the "true" parameters (or posterior mode, if you prefer) and, in turn, to compare this vis-á-vis extant estimation techniques. With this in mind, I construct a simple Monte Carlo experiment to examine the properties of this estimator.

The sample size for this analysis will be set to $1,000$.[11] For simplicity, I assume the covariates for the latent opinion, decisiveness, and the random intercept are the same. Specifically, for the latent opinion, ϑ_i,

$$\vartheta_i = \beta_0^{\vartheta} + \beta_1^{\vartheta} x_1^{\vartheta} + \varepsilon_i, \tag{23}$$

where $\varepsilon_i \sim \mathcal{N}(0,1)$. The first term is simply a constant intercept. The variable x_1^{ϑ} is simply a collection of $1,000$ draws from the Standard Normal Distribution. This variable will be the same for the decisiveness and individual-saliency effects, so I abbreviate it henceforth as x_1.

The decisiveness equation is derived similarly:

$$\delta_i = \beta_0^{\delta} + \beta_1^{\delta} x_1 + \varepsilon_i. \tag{24}$$

Saliency is a bit trickier due to the hierarchy. At the highest level, it is Normal with mean equal to $\eta_i + X_i^{\psi}\beta^{\psi}$ and variance 1. The individual saliency intercepts, η_i, are in turn Normal with mean $X_i\beta^{\eta}$ and variance σ_{η}^2. The full draw for saliency can be written as

$$\begin{aligned} \psi_i &= X_i^{\eta}\beta^{\eta} + X_i^{\psi}\beta^{\psi} + \varepsilon_i + \varepsilon_i' \\ &= \beta_0^{\eta} + \beta_1^{\eta} x_1 + \beta_1^{\psi} x_1 + \beta_2^{\psi} x_2 + \varepsilon_i + \varepsilon_i' \end{aligned} \tag{25}$$

[11] The choice of $1,000$ is in a certain sense arbitrary, as it can modified without noticeable differences in results. My choice to use this level hinges on two points. First, smaller samples tend to lead the ordered probit model to artificially "fall apart," as the NA values become a greater percentage of the data. Second, very large samples (e.g., 5,000 or 10,000) tend to mute the effects of missing values, except when the missingness is heavily systematic. Since my goal is for this method to be employed by those who analyze survey data, the choice of $1,000$ fits squarely within the range of usual sample sizes, making this Monte Carlo analysis more sensible.

where ε_i' is Normal with mean 0 and variance σ_η^2. x_1 is the same as in the previous models, but now I introduce x_2, which is just a uniform random variable on the range -2 to 4.

All parameters, β's and σ^2's, are set to 1 and the cutpoints are set to

$$c = (-\infty, 0, 1.0, 1.5, 2.0, 2.5, 3.0, \infty).$$

These cutpoints are then used to define the observed ordinal y's. Finally, individuals whose saliency draws were less than zero *or* whose decisiveness draws within the region of indifference were less than zero had their ordinal y replaced with an missing value.

A histogram of the non-NA y's is presented in Figure 3. This figure shows a high clustering at the low end of the scale, with a fair degree of uniformity to the right of the first two categories. A *prima facie* analysis of this data would suggest that respondents tend to be on the liberal end of the scale, provided we interpret y as an ideological scale. At this point, one would usually use ordered probit to analyze the predicted probabilities of falling into one of these bins.

Figure 3 about here.

The major challenge to any subsequent analysis is the presence of missing values. Based on the parameters chosen, the number of NA's found was 306 out of the total sample of $1,000$ observations. These include cases due to lack of saliency and lack of decisiveness when in the region of indifference. Note however that the *latent* opinion is unaffected by missingness. This important feature allows the results from the hierarchical model to be compared directly to the ordered probit model. The ordered probit is derived in precisely the same manner as the latent ϑ_i's. Since that model cannot accommodate the NA's, however, they must be deleted from the analysis. The deletion of these cases *should* bias the estimates of the latent opinion slopes, β^θ.

To examine this further, I use Martin and Quinn's `MCMCpack` (Martin and Quinn 2007) package in `R` to estimate the Bayesian ordered probit model, as well as my own C++ code, employing the Scythe Statistical Library (Martin and Quinn 2001) to estimate the hierarchical model described above.[12] In both cases, priors were identical for the β's and cutpoints (see Section 2). Starting values were Standard Normal random draws for all parameters except the cutpoints and σ_η^2 (for the hierarchical model). The cutpoints were evenly spaced over the interval 0 to 5 and the variance was drawn from an Inverse Gamma Distribution with shape and scale of 1. For each case, 100,000 iterations were used in the analysis. The first half were discarded as burn-in. Due to storage limitations with the hierarchical model, every 20^{th} iteration was saved. Three chains were examined for each and standard convergence diagnostics were employed to check convergence (i.e., Gelman and Rubin's [1992] $\hat{R}$ and Geweke's [1992] diagnostic). Both models were found to have converged.

While the coefficient estimates are not that different for the two models, the divergence in substantive effects (i.e., predicted probabilities) is quite significant.[13] In the hierarchical model, for a given specification of the regressors, predicted probabilities are given by equations (10), (11), and (12). For the ordered probit, predicted probabilities are given by

$$Pr(y_i = q) = \Phi(c_q - \mu_i^{\vartheta_i}) - \Phi(c_{q-1} - \mu_i^{\vartheta_i}). \tag{26}$$

Note that for non-NA categories, the degree of divergence between the ordered probit and the hierarchical model depends on saliency and decisiveness. If there is a variable that affects both saliency and latent opinion, it in natural to think that our inferences in terms of predicted probabilities will differ. To see this

[12] For the C++ code, I am indebted to Martin and Quinn, as their Scythe Statistical Library (2001) greatly assisted in the writing and running of code.

[13] As the posteriors are not too different, presenting the relevant plots is not very useful. Nonetheless, they are available from the author upon request.

analytically, consider a response that is a non-NA, non-indifference region value. The divergence between predictions is given by dividing equation (11) by equation (25), which yields $1 - \Phi(-\mu_i^{\psi})$. If the mean is very large (and positive), the Normal cdf approaches zero and the the two models should be approximately equal. However, as the mean gets very small (and negative), the cdf approaches one and the models are maximally divergent.

As a concrete example of this divergence, set x_2 and the random intercept, η, at its mean value and allow x_1 to vary along its domain. This allows calculations of the predicted probabilities for both models. A plot of the predicted probability of the respondent choosing category 1 in each of the models is found in Figure 4. As x_1 increases, the ordered probit suggests the probability of being in this category should decrease monotonically from 1 to 0. The hierarchical model suggests a different story, however. At low values of x_1, the predicted probability is 0. The probability then increases until about -1, where it begins to decrease and match the ordered probit results. This non-monotonic result leads to starkly different inferences than the ordered probit model.

To make this contrast even clearer, suppose that y represents a seven-point ideology scale and x_1 is the income of the respondent. The ordered probit predictions suggest that very poor respondents are very likely to pick the extremely liberal category. As income increases, this probability goes to zero. The hierarchical model presents a different picture, whereby the very rich and very poor have a low predicted probability of choosing the extremely liberal category. The lower-middle class individuals are the ones with the highest probability of choosing this category.

Figure 4 about here.

The divergence in inference is exacerbated when one considers that the saliency intercept, η, was held at its mean in deriving the hierarchical probabilities. If this value is allowed to take on its minimum value, the gap in predictions is even

wider. Figure 5 shows that the predicted probabilities are essentially zero for choosing $y = 1$ for all values of x_1. Thus, the two models give vastly different predictions for $x_1 \in [-3, 1)$.

Figure 5 about here.

To obtain a more general view of how varying both the random saliency intercept and x_1 affects the predicted probabilities, I present two additional plots. Figure 6 shows the probabilities for the ordered probit model. Not surprisingly, varying η has no effect on the probabilities, as this term does not affect the probability model. Figure 7 shows the probabilities for the hierarchical model. There is noticeable variation in both variables. Comparing this plot with Figure 6 demonstrates that the models agree only when the intercept η is extremely high or when x_1 is extremely low. Thus, other than in these extreme circumstances, the two models produce widely different inferences.

Figure 6 about here.

Figure 7 about here.

Appendix B: Figures

Figure 1: The hierarchical model

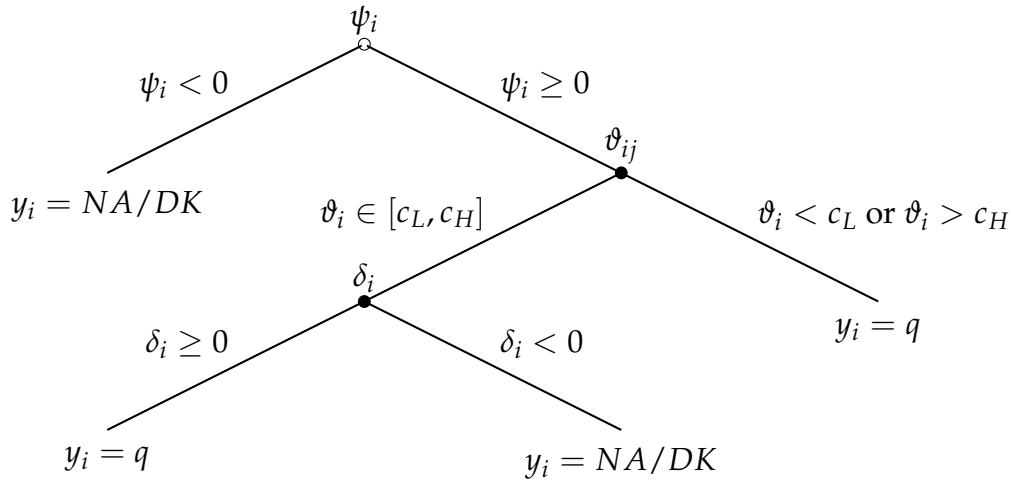

Figure 2: The hierarchical model with probabilities

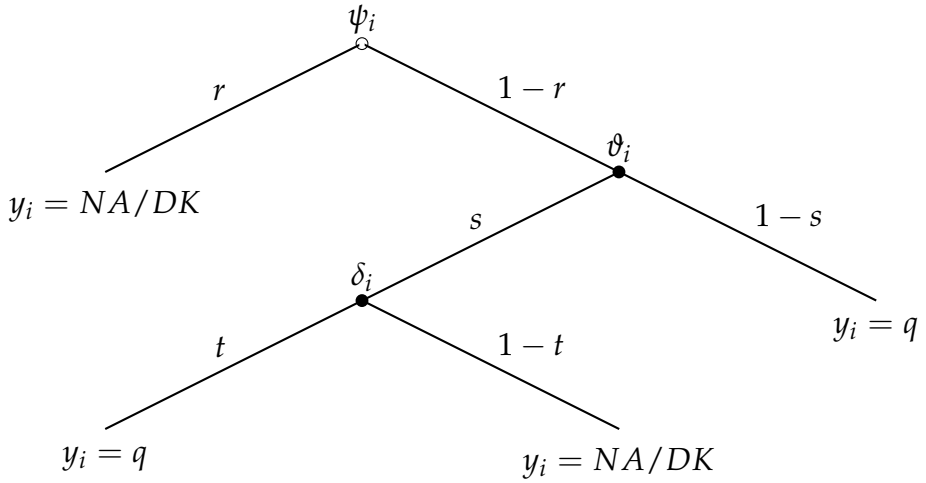

Figure 3: Ordinal y-values for the Monte Carlo experiment

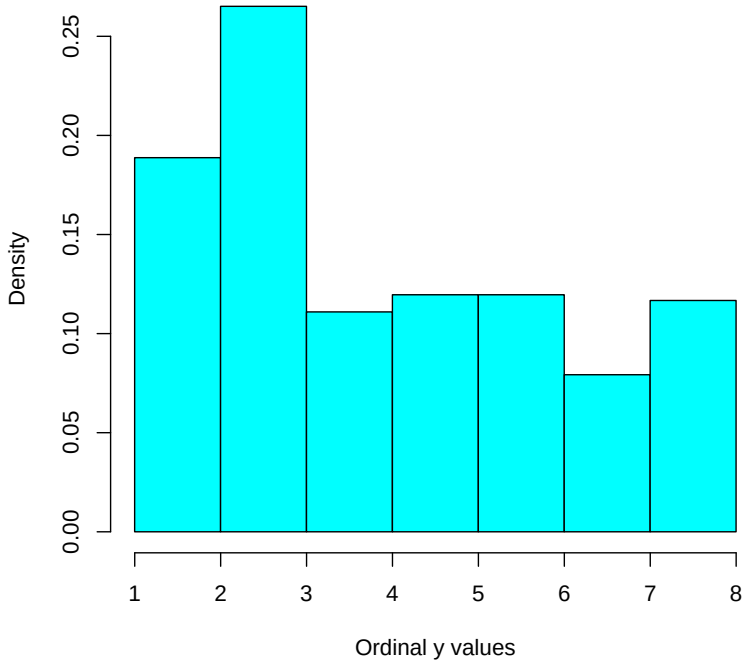

Figure 4: When Ordered Probit Goes Bad

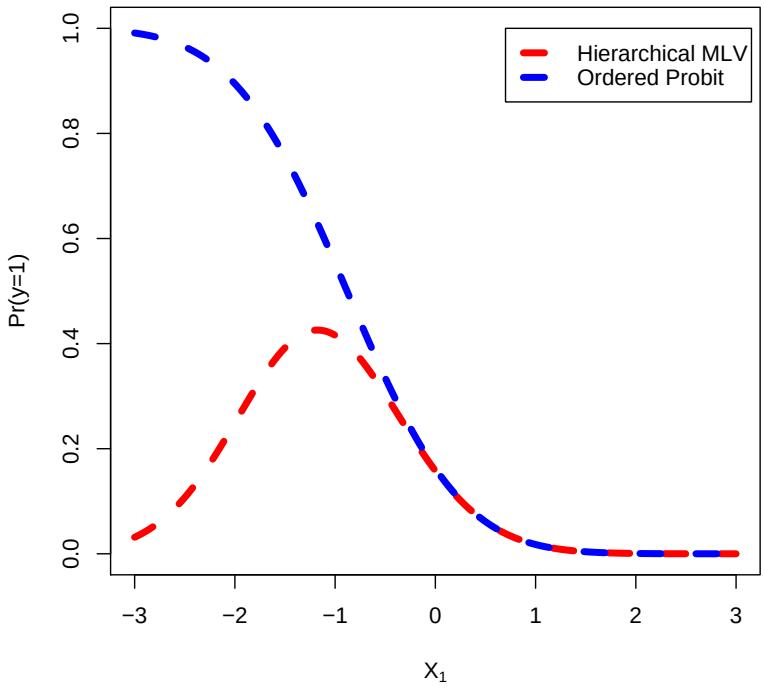

Figure 5: When Ordered Probit Gets Worse

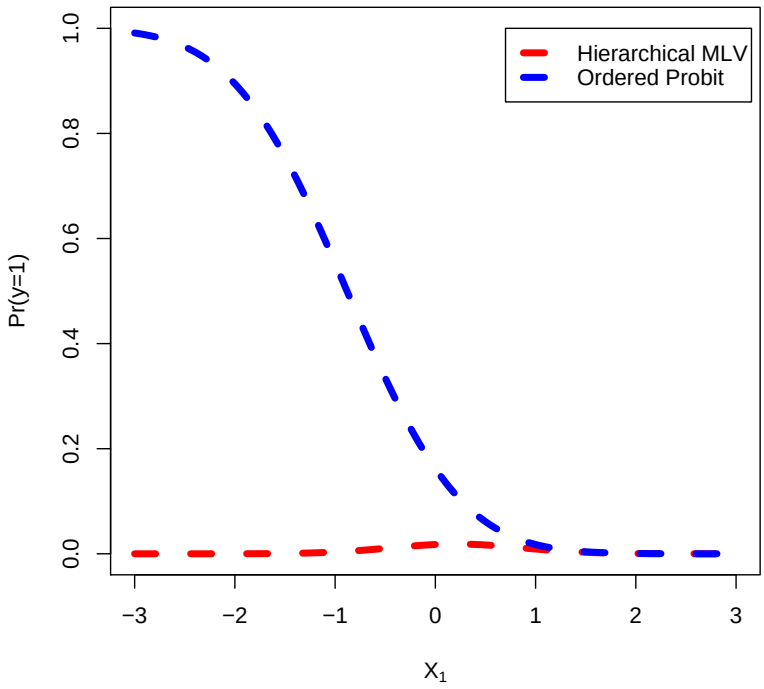

Figure 6: Ordered Probit Probabilities of $y = 1$ varying both η and x_1

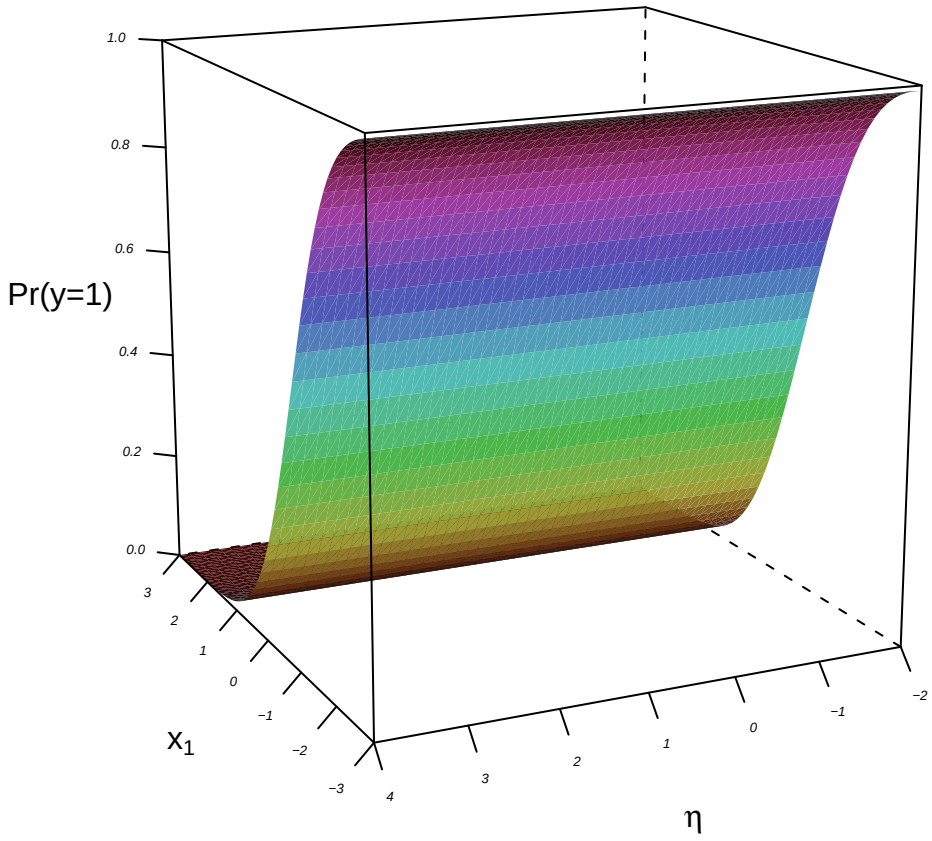

Figure 7: Hierarchical Model Probabilities of $y = 1$ varying both η and x_1

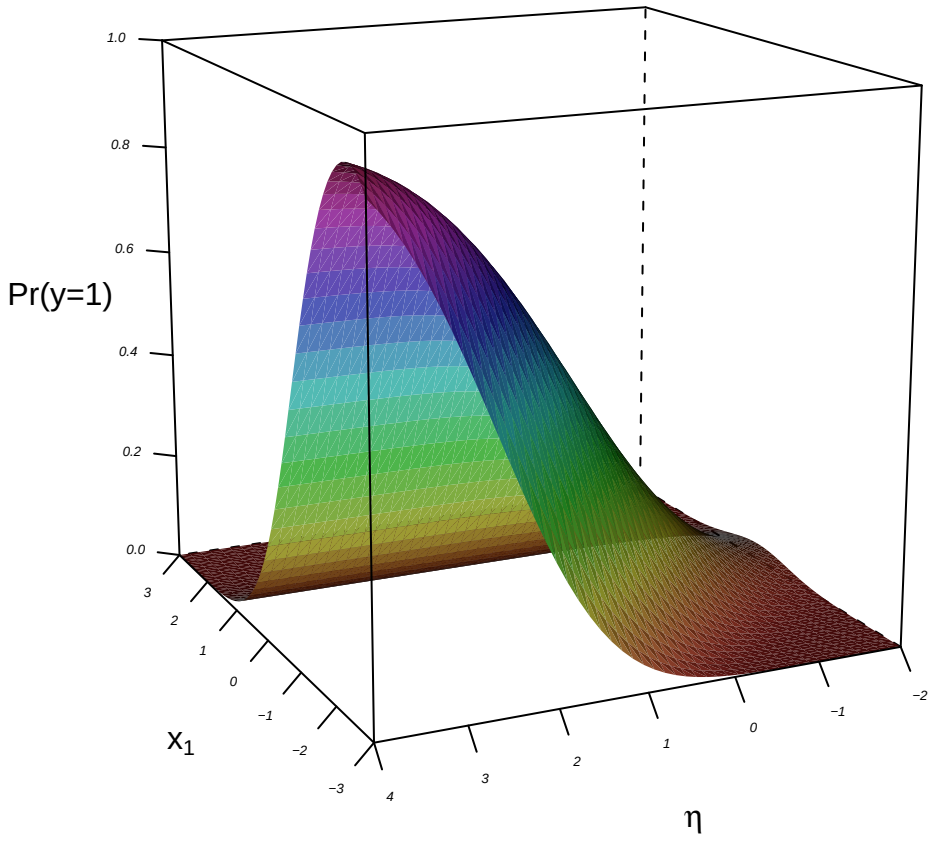

Table 1: Placement of Republican House Candidates in the Hierarchical Model

		Posterior mean	Standard deviation
Opinion equation			
	Intercept	0.6106	0.3602
	Age	0.0065	0.0035
	Black	0.3949	0.2275
	Male	−0.1113	0.1084
	Education	−0.0088	0.0243
	Democrat	**0.3010**	0.1377
	Republican	**0.2795**	0.1347
	Extremism	**0.2863**	0.0626
	Info	**0.1047**	0.0200
Decisiveness equation			
	Intercept	0.2044	1.7298
	Age	0.0890	0.1031
	Black	−0.0213	1.8396
	Male	−0.1128	1.7891
	Education	**0.9379**	0.4944
	Democrat	0.0973	1.6982
	Republican	−0.5199	1.5783
	Extremism	0.1520	1.5411
	Info	−0.4896	0.6368
Saliency intercept equation			
	Intercept	−**1.0390**	0.4223
	Age	**0.0066**	0.0043
	Black	−0.2216	0.2429
	Male	0.1834	0.1415
	Education	0.0144	0.0290
	Democrat	−0.1802	0.1626
	Republican	0.0371	0.1886
Saliency equation			
	Info	**0.0352**	0.0225
	Extremism	**0.1449**	0.0752

Table 2: Placement of Rep. House Candidates—Ordered Probit Model with List-wise Deletion

	Posterior mean	Standard deviation
Intercept	**0.9022**	0.3656
Age	**0.0075**	0.0038
Black	**0.4738**	0.2497
Male	−0.1167	0.1257
Education	−0.0071	0.0279
Democrat	0.2781	0.1540
Republican	**0.3139**	0.1551
Extremism	**0.3038**	0.0703
Info	**0.1133**	0.0217

Table 3: Placement of Rep. House Candidates—Ordered Probit Model with MI

	Posterior mean	Standard deviation
Intercept	**1.3430**	0.2689
Age	**0.0077**	0.0024
Black	**0.3090**	0.1365
Male	−0.0414	0.0772
Education	0.0330	0.0173
Democrat	**0.2160**	0.0921
Republican	**0.3449**	0.0971
Extremism	**0.2921**	0.0436
Info	**0.0457**	0.0131

References

[1] Albert, James and Siddhartha Chib. 1993. "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association* 88, 422: 669-679.

[2] Althaus, Scott. 1998. "Information Effects in Collective Preferences." *American Political Science Review* 92: 545-558.

[3] Bartels, Larry. 1998. "Panel Attrition and Panel Conditioning in American National Election Studies." Paper presented at the 1998 meetings of the Society for Political Methodology, San Diego.

[4] Bradlow, Eric and Alan Zaslavsky. 1999. "A hierarchical latent variable model for ordinal data from a customer satisfaction survey with 'no answer' responses." *Journal of the American Statistical Association* 94, 445: 43-52.

[5] Berinsky, Adam. 1999. "The Two Faces of Public Opinion." *American Journal of Political Science* 43: 1209-1230.

[6] Brehm, John. 1993. *The Phantom Respondents: Opinion Surveys and Political Representation.* Ann Arbor: University of Michigan Press.

[7] Clinton, Joshua, Simon Jackman, and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98: 355-370.

[8] Conover, Pamela Johnston and Stanley Feldman. 1989. "Candidate Perception in an Ambiguous World: Campaigns, Cues, and Inference Processes, " *American Journal of Political Science* 33: 912-940.

[9] Converse, Philip. 1964. "The nature of belief systems in mass publics." In D. E. Apter (ed.), *Ideology and discontent* (pp. 206-261). New York: Free Press of Glencoe.

[10] Delli Carpini, Michael X. and Scott Keeter. 1993. "Measuring Political Knowledge: Putting First Things First." *American Journal of Political Science* 37: 1179-1206.

[11] —. 1997. *What Americans Know about Politics and Why It Matters*. New Haven: Yale University Press.

[12] Feldman, Stanley and Pamela Johnston Conover. 1983. "Candidates, Issues, and Voters: The Role of Inference in Political Perception." *Journal of Politics* 45: 810-839.

[13] Gelman, Andrew and David Rubin. 1992. "Inference from iterative simulation using multiple sequences." *Statistical Science* 7: 457-511.

[14] Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2004. *Bayesian Data Analysis*. 2nd ed. New York: Chapman and Hall.

[15] Gelman, Andrew, Gary King, and Chuanhai Lin. 1999. "Not Asked and Not Answered: Multiple Imputation for Multiple Surveys." *Journal of the American Statistical Association* 93 (September): 84657; with comments by John Brehm, David R. Judkins, Robert L. Santos, and Joseph B. Kadane, and rejoinder by Gelman, King, and Liu, pp. 86974.

[16] Geweke, John. 1992. "Evaluating the accuracy of sampling-based approaches to calculating posterior moments." In *Bayesian Statistics 4* (ed JM Bernado, JO Berger, AP Dawid and AFM Smith). Clarendon Press, Oxford, UK.

[17] Heckman, James. 1976. "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables, and Simple Estimator for Such Models." *Annals of Economic and Social Measurement* 5: 47592.

[18] King, Gary, James Honaker, Anne Joseph and Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." *American Political Science Review* 95: 49-69.

[19] King, Gary, James Honaker, and Matthew Blackwell. 2007. "Amelia II: A Program for Missing Data." http://gking.harvard.edu/amelia.

[20] Koch, Jeffrey. 2000. "Do Citizens Apply Gender Stereotypes to Infer Candidates' Ideological Orientations?" *The Journal of Politics* 62 (2), 414429.

[21] Martin, Andrew and Kevin Quinn. 2001. "MCMCpack, Release 0.9-3." http://mcmcpack.wustl.edu.

[22] —. 2007. "Scythe Statistical Library, Release 0.1." http://scythe.wustl.edu.

[23] McDermott, Monika. 1997. "Voting Cues in Low-Information Elections: Candidate Gender as a Social Information Variable in Contemporary United States Elections." *American Journal of Political Science* 41: 270-283.

[24] Miller, Warren, Donald R. Kinder, Steven J. Rosenstone, and the National Election Studies. 1993. *American National Election Study, 1992: Pre- andPost-Election Survey* [Computer file]. Ann Arbor, Mich.: University of Michigan, Center for Political Studies/Inter-university Consortium for Political and SocialResearch. ICPSR Study No.6067.

[25] Jackman, Simon. 2000. "Estimation and Inference via Bayesian Simulation: An Introduction to Markov Chain Monte Carlo." *American Journal of Political Science* 44: 375-404.

[26] Peress, Michael. 2007. "Correcting for Survey Nonresponse using Variable Response Propensity." Unpublished.

[27] Poole, Keith and Howard Rosenthal. 1997. *Congress: A Political-Economic History of Roll Call Voting*. New York: Oxford University Press.

[28] Rahn, Wendy. 1993. "The Role of Partisan Stereotypes in Information Processing about Political Candidates ." *American Journal of Political Science* 37: 472-496.

[29] Rubin, Donald. 1976. "Inference and missing data." *Biometrika* 63(3):581-592.

[30] —. 1977. "Formalizing Subjective Notions about the Effect of Nonrespondents in Sample Surveys." *Journal of the American Statistical Association* 72 (September): 538-543.

[31] —. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

[32] Sigelman, Lee, Carol Sigelman, Barbara Walkosz, and Michael Nitz. 1995. "Black Candidates, White Voters: Understanding Racial Bias in Political Perceptions." *American Journal of Political Science* 39: 243-265.

[33] Tanner, Martin and Wing Hung Wong. 1987. "The Calculation of Posterior Distributions by Data Augmentation." *Journal of the American Statistical Association* 82, 398: 528-540.

[34] Zaller, John. 1992. *The Nature and Origins of Mass Opinion*. New York: Cambridge University Press.