Bayesian Partially Ordered Multinomial Probit/Logit Model with an Application to Course Redesign

Michael D. Sonksen * [1], Xueqin Wang† [1], and Kristin Umland [1]

[1]The University of New Mexico, Albuquerque, NM 87131, USA

Abstract

Large entry-level courses are commonplace at public 2- and 4-year institutions of higher education (IHEs) across the United States. Low pass rates in these entry-level courses, coupled with tight budgets, have put pressure on IHEs to look for ways to teach more students more effectively at a lower cost. Efforts to improve student outcomes in such courses are often called "course redesigns." The difficulty arises in trying to determine the impact of a particular course redesign; true random-controlled trials are expensive and time-consuming, and few IHEs have the resources or patience to implement them. As a result, almost all evaluations of efforts to improve student success at scale rely on observational studies. At the same time, standard multi-level models may be inadequate to extract meaningful information from the complex and messy sets of student data available to evaluators because they throw away information by treating all passing grades equally. We propose a new Bayesian approach that keeps all grading information: a Partially Ordered Multinomial Probit Model with random effects fit using a Markov Chain Monte Carlo algorithm, and a Logit model can be fit with importance sampling. Simulation studies show that the Bayesian Partially Ordered Probit/Logit Models work well, and the parameter estimation is precise in large samples. We also compared this model with standard models considering Mean Squared Error and the area under the Receiver Operating Characteristic (ROC) curve.

Key Words: Markov Chain Monte Carlo, latent variable models, multilevel logistic regression, online learning, model diagnostics

1. Introduction

Multilevel models (i.e., random effect or hierarchical models) have become a popular choice in applied statistics. Raudenbush (1988) reviewed statistical methods in educational statistics and concluded that multilevel linear models were a prominent theme in modeling education data. Kaplan and Elliott (1997) summarized that the application of multilevel linear regression methods had resulted in extraordinary advances in school process research since multilevel linear models account for the clustered sampling schemes in education research. Their popularity has only increased over time.

Common models used for categorical response data are probit/logit models (Wong and Mason, 1985; Stiratelli et al., 1984) and ordered probit/logit models (Schaafsma and Osoba, 1994; Fullerton, 2009; Hedeker, 2008). These approaches enable the researcher to model binary or ordinally scaled dependent variables with one or more independent variables. The response data for ordered probit/logit models has to be ordered with equal distance between each point, for example, a 5-point Likert scale. All of these models can be easily fit in software packages such as SAS (PROC GLIMMIX, NLMIXED, and LOGIT among others) and R ($nlme$ and $polr$ libraries among others). Maximum likelihood estimation is primarily used by these procedures.

*sonksen@stat.unm.edu

†xueqin@unm.edu

However, we often have data that does not fit into one of these scenarios. For example, assume that we observe students' letter grade with possible values: A, B, C, D, F and CR (Pass with Credit), NC (Fail with No Credit). This variable is partially ordered because, while an A is clearly greater than a D, a CR is not necessarily greater or less than an A, B, or C. In order to fit a multilevel probit/logit model to this type of data, we must collapse all passing grades together and all failing grades together. Similarly, to fit an ordered probit/logit model we must only consider the fully ordered letter grade (A-F). Both of these approaches result in a loss of information by either trucation or exclusion.

For computational reasons, we consider Bayesian models in this work. Albert and Chib (1993) first brought a Bayesian method using a latent variable to model dichotomous and polychotomous ordered and unordered response data. This paper expands the model of Albert and Chib (1993) to partially ordered response data with random effects in a fully Bayesian approach. We also extend the results to a logit model using the idea of O'Brien and Dunson (2004). We applied this model to an example of a course redesign evaluation at a large public university, which we will call the University, in the southwestern United States.

1.1 Motivating Example

There are a number of large-enrollment, entry-level courses at the University that students are placed into, and must pass, in order to take subsequent courses or satisfy University core curriculum requirements. The course undergoing a redesign that we evaluated in this study, which we will call the Course, enrolls more than 2500 students each year and has a historical failure rate that is over 50%. Prior to the Fall Semester 2012, all sections of the Course had 60 students in which each instructor chose the method of instruction. Historically, most instructors chose a traditional lecture model, although some instructors had tried different approaches to improve student performance. No formal evaluation on any of these efforts was ever undertaken.

In the Fall Semester of 2012, the University ran two pilot sections of the Course that used computer-based instruction, which we will call the Redesigned Course. An important component of the redesign for this analysis is that students were required to show mastery of individual topics before moving on to the next topic, and students who did not show mastery of all required topics by the end of the term were given a grade of Incomplete and were allowed to continue working without re-enrolling in the course the following semester. A total of 216 students (about 16.5%) enrolled in the pilot of the Redesigned Course, and 1,092 students enrolled in a standard section of the Course. The research team first set out to answer the following questions:

1. What is the impact of the Course redesign on student success?

   Available resources and time pressure did not permit the University to evaluate the pilot Redesigned Course using a true experimental paradigm. Thus, the only available option was to do an observational study that models student success using student and instructor covariates to predict students' course grades. Additionally, a large percentage of students enrolled in the pilot course (54.6%) received a grade of Incomplete, and so it was not possible to determine whether they would eventually be successful or not at the time of the study. In the process of trying to answer the first question, the research team came to a second question:

2. What are good model structures for analyzing this kind of data?

1.2 Data Description

The data was obtained from the University's grade repository. The original data consists of information for 1308 students enrolled in the Course during the Fall Semester of 2012. We dropped variables that were not relevant to our research questions. The cleaned and de-identified data contains the following variables: a math achievement score that combines normalized SAT and ACT scores (different students took different standardized tests, so it was necessary to combine them into a single variable), course credit-hour load for Fall 2012, Fall 2012 semester GPA (excluding grades from the course), high school attended, high school graduation year, gender, race/ethnicity, and Course instructor. The instructor variable was treated as a random effect in all models described here because we are interested in comparing students' performance in the Redesigned Course to that of students in the standard version of the Course, and both sections of the Redesigned Course were taught by the same instructor in the Fall Semester of 2012. Our response variable is the students' letter grades (Y_i): A, B, C, D (with +/-), F , CR (Passing), NC (Failing), W (Withdrawal), I (Incomplete), AUD (Auditing). The remaining variables were assumed to be fixed independent covariates in all models considered in this work.

Choosing a model for all of this data is complicated because the possible values for Y_i are letter grades that are not totally ordered. The data is discrete with more than two categories, but has additional structure than a standard multinomial likelihood model can adequately handle. Many authors who have contributed to this literature discarded information in the data in order to use a standard model such as a multilevel (possibly ordered) probit/logit models.

1.3 Limitations of Common Models

One option for modeling partially ordered data such as letter grades is a multilevel logistic model (Wong and Mason, 1985; Gilmour et al., 1985; Gelman and Hill, 2007), in which,

$$\begin{array}{rcll} Y_i|p_i & \overset{ind}{\sim} & Bern(p_i) & \\ logit(p_i) & = & \boldsymbol{X}_i{}'\boldsymbol{\beta} + \boldsymbol{Z}_i{}'\boldsymbol{\tau} & i = 1, 2, \ldots, n. \\ \tau_j|\sigma^2_\tau & \sim & N(0, \sigma^2_\tau) & j = 1, 2, \ldots, J. \end{array}$$

Y_i is a binary response variable. $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are fixed and random design vectors for student i; $\boldsymbol{X}_i$ contains the student's demographic and achievement information and $\boldsymbol{Z}_i$ indicates which instructor the student had. $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)'$ is a vector of unknown fixed effects and $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_J)'$ is the vector of unknown instructor random effects. σ^2_τ is the variance of the random effect τ, and n is the number of students, while k is the number of covariates and J is the number of instructors. A probit model can be used by simply replacing the logit function with the inverse standard normal cdf.

With this model, we are forced to make our letter grade response variable into a binary variable (pass or fail). This means that the passing letter grades (A+, A, A- ,B+, B, B-, C+, C, CR) are treated equally. Similarly, the failing letter grades (C-, D+, D, D-, F, NC) are also treated equally. It is important to note that this model cannot describe student success precisely. Suppose there are two classes of 20 students each; in one class, 10 students passed the course with an A and 10 failed with a F, while 10 students in the other class passed the course with a C and 10 failed with a F. The pass rate is the same between the two classes, 50%, but the student performance is actually different. Another issue in our data is that many students in the Redesigned Course (54.6%) received an Incomplete at the end of the term. How should we interpret these Incompletes: as a passing letter grade? As a failing letter grade? These students had not yet completed all of the course topics, but they were

still working. Both options are extreme because this "go at your own pace" course can be completed over several semesters; there is simply not enough information available to make this determination *a priori*. We could leave these students out of the study, but eliminating over half of our treatment group would be highly problematic.

Another option for modeling this type of data is to use an ordered multinomial logit model, which enables us to model ordinally scaled or unordered nominal dependent variables with independent variables. For early descriptions of ordered logit models see McKelvey and Zavoina (1975), McCullagh (1980), and Winship and Mare (1984). An ordered multinomial logit model is an extension of the above logistic models, but the response data can have more than two categories. Ordered multinomial logit models are widely used in economics, education and psychology. Common examples of ordered categories are the Likert scale "strongly dissatisfied, dissatisfied, neutral, satisfied, strongly satisfied" or frequency "Always, frequently, sometimes, rarely, never", income "$0-10K/year, 10-20K/year, 20-30K/year, 30-60K/year, >60K/year$". With this model, the dependent variable has to include ordered cases with equal distance between each point. While some of our response data are ordered (A, B, C, D (with +/-s), F), some are not ordered and overlap (CR, NC, W, I, AUD). Note that NC, W, I are not necessarily better than D or F, and CR is not necessarily better than A, B, or C. Thus, only part of our data qualifies for the ordered multinomial logit model, and we would have to remove observations with responses CR, NC, W, I, AUD to use these models. This would result in a loss of information. Again, a probit version of this model can be easily considered by changing the link function.

In this paper, we propose two new models, the Bayesian Partially Ordered Multinomial Probit and Logit Models, which can accommodate data which have some ordering but are not fully ordered. This rest of this paper is organized as follows. Section 2 describes the new model we employed in this study. Section 3 presents computational strategies for fitting the models. The advantages of utilizing the new models are summarized in Section 4. We conduct a simulation study in Section 5. Section 6 applies the new models to our Redesigned Course evaluation at the University.

2. A Description of Our Proposed Models

Assume that each student has a latent grade, S_i, a numerical score for student i. If we knew the score, a reasonable model would be a hierarchical linear model:

$$S_i = \boldsymbol{X}_i'\boldsymbol{\beta} + \boldsymbol{Z}_i'\boldsymbol{\tau} + \epsilon_i \tag{1}$$

With

$$\begin{aligned} \epsilon_i|\sigma_\epsilon^2 &\overset{iid}{\sim} N(0, \sigma_\epsilon^2) \\ \tau_j|\sigma_\tau^2 &\overset{iid}{\sim} N(\mu_\tau, \sigma_\tau^2) \end{aligned}$$

for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, J$. Where, $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are the fixed and random design vectors, respectively, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)'$ is a vector of unknown fixed effects, and $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_J)'$ is the vector of unknown instructor random effects. We denote the sample size by n, the number of fixed effects by k, and the number of instructors by J.

In our example, we do not know S_i. However, our observed letter grade Y_i gives us information on the latent score S_i. For example, if student i passed, we know that $S_i > 73$ if 73 is the cut-off point for passing the course.

Albert and Chib (1993) examined the situation where Y_i is either a pass or fail with a Bayesian probit model. Note that, if we assume $\epsilon_i \overset{iid}{\sim} N(0,1)$

$$P(S_i > 73|\boldsymbol{\beta}, \boldsymbol{\tau}) = 1 - \Phi(73 - X_i\beta - Z_i\tau) = \Phi(-(73 - X_i\beta - Z_i\tau)).$$

Which gives us a probit-style probability. For the case of ordered Y_i (for example A, B, C, D, F), Albert and Chib (1993) assumed a latent continuous random variable distributed as $N(\boldsymbol{X\beta}, 1)$, and the observed categorical response Y_i = j if $\gamma_{j-1} < S_i \leq \gamma_j$, where the bin boundaries $\gamma_1, ..., \gamma_{J-1}$ are unknown, and $\gamma_0 = -\infty, \gamma_J = \infty$ are defined. So $Si|\beta, \gamma, Y$ is truncated normal. This forms a Bayesian ordered probit model.

In our case, including CR, NC, I, W, AUD as possible letter grades, we can only partially order the responses. However, each possible grade does yield a range of possible values of S_i, for example, we can view $Y_i = A+ \Rightarrow 97 \leq S_i, \quad Y_i = A- \Rightarrow 93 \leq S_i < 97, \ldots Y_i = F \Rightarrow 0 \leq S_i < 60$. For the unorderable grades, we assume that $Y_i = CR \Rightarrow 73 \leq S_i$, $Y_i = NC \Rightarrow 0 \leq S_i < 73$, $Y_i = W \Rightarrow 0 \leq S_i < 73$, $Y_i = I \Rightarrow 0 \leq S_i \leq 100$, and $Y_i = AUD \Rightarrow 0 \leq S_i \leq 100$. This assumes that we have little knowledge what the course score of a student receiving an Incomplete, Withdrawal, or Audit was at the time he or she left the course. See Table 1 for these ranges based on the Course grading scale. The key purpose of this formulation is that we know the conditional distribution of the latent data S_i conditional on the parameters $(\boldsymbol{\beta}, \boldsymbol{\tau}, \sigma_\tau^2, \sigma_\epsilon^2)$ and letter grade (Y_i) has a truncated normal distribution:

$$S_i|\boldsymbol{\beta}, \sigma_\epsilon^2, \boldsymbol{\tau}, Y_i, \sigma_\tau^2 \overset{ind}{\sim} TN(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\tau}, \boldsymbol{\sigma_\epsilon^2}, L_{Y_i}, U_{Y_i}).$$

Let L_{Y_i} denote the lower bound and U_{Y_i} denote the upper bound of S_i given student i received letter grade Y_i. Then, S_i is truncated on the boundaries of each letter scale, and we know that $S_i \in (L_{Y_i}, U_{Y_i})$. Table 1 displays the boundaries of course scores corresponding to each letter grade. If a student earned an A+, his course score is between 97 and 100, an A is between 93 and 97, ..., and an F is between 0 to 60. We used 73 as the cutoff passing score because a C- is not considered a passing grade for this Course, so the score for CR is between 73 and 100, while NC is between 0 and 73. A "W" is treated as a failure for this work. "I" and "AUD" do not give us much information about students' score, so "I" and "AUD" can take any value between 0 and 100.

Letter Grade	L_{Y_i}	U_{Y_i}	Letter Grade	L_{Y_i}	U_{Y_i}
A+	97	100	D+	67	70
A	93	97	D	63	67
A-	90	93	D-	60	63
B+	87	90	F	0	60
B	83	87	CR	73	100
B-	80	83	NC	0	73
C+	77	80	W	0	73
C	73	77	I	0	100
C-	70	73	AUD	0	100

Table 1: The range of course scores corresponding to each letter grade L_{Y_i} is the lower limit score of a letter grade, and U_{Y_i} is the upper limit of the letter grade.

If a logit style model is desired, O'Brien and Dunson (2004) showed that a simple change of the distribution of ϵ_i results in a logit marginal distribution. The logit link can be implemented in the model listed in Equation (1) by changing the distribution of $\boldsymbol{\epsilon}$ to

$\epsilon_i \overset{ind}{\sim} logistic(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\tau}, \sigma^2_\epsilon)$. The predicted probability with the logit link is easier to interpret than with the probit link.

The key point to take away from this model is that S_i is unknown, and we only observe the partially ordered categorical variable Y_i. However, the letter grade a student receives does provide information on the possible value of S_i. With this latent variable representation, we can not only model ordered and unordered multinomial response data, but also partially ordered categorical outcomes. The knowledge from the partial ordering is also preserved. Latently, our model has exactly the form of a linear mixed model with random effects. We included random effects in the model because the grades of students taught by the same instructor may be correlated.

In this work, we utilize a Bayesian approach for several reasons. First, the computation is easier for predictions. Sampling from the posterior allows us to make predictions while taking into account parameter uncertainty. Second, Bayesian analysis can easily incorporate truncated distributions and also allows us to use subjective information in our priors.

To complete a Bayesian analysis, we need prior distributions for $\boldsymbol{\beta}$, $\boldsymbol{\tau}$, μ_τ, σ^2_τ, and σ^2_ϵ. For computational convenience, we assumed an improper uniform prior for $\boldsymbol{\beta}$, a normal prior for the instructor random effects $\boldsymbol{\tau}$ (conditional on hyperparameters), an improper uniform prior for the average instructor random effect μ_τ, and inverse gamma priors for the variance of the instructor random effects σ^2_τ and the latent score error variance σ^2_ϵ. Symbolicly:

$$\begin{aligned}
\pi(\boldsymbol{\beta}) &\sim 1 \\
\tau_j|\mu_\tau, \sigma^2_\tau &\overset{iid}{\sim} N(\mu_\tau, \sigma^2_\tau) \text{ for } j = 1, 2, \ldots, J \\
\pi(\mu_\tau) &\propto 1 \\
\sigma^2_\tau &\overset{iid}{\sim} IG(a_\tau, b_\tau) \\
\sigma^2_\epsilon &\overset{iid}{\sim} IG(a_\epsilon, b_\epsilon)
\end{aligned}$$

where, $a_\tau, b_\tau, a_\epsilon, b_\epsilon$ are fixed constants. In practice, these may be chosen to reflect subjective knowledge or prior ignorance. Typical conditional independence is assumed, meaning:

$$\begin{aligned}
\pi(\boldsymbol{\beta}, \boldsymbol{\tau}, \sigma^2_\tau, \sigma^2_\epsilon) &= \pi(\boldsymbol{\beta}|\boldsymbol{\tau}, \mu_\tau, \sigma^2_\tau, \sigma^2_\epsilon)\pi(\boldsymbol{\tau}|\mu_\tau, \sigma^2_\tau, \sigma^2_\epsilon)\pi(\mu_\tau|\sigma^2_\tau, \sigma^2_\epsilon)\pi(\sigma^2_\tau|\sigma^2_\epsilon)\pi(\sigma^2_\epsilon) \\
&= \pi(\boldsymbol{\beta})\pi(\boldsymbol{\tau}|\mu_\tau, \sigma^2_\tau)\pi(\mu_\tau)\pi(\sigma^2_\tau)\pi(\sigma^2_\epsilon)
\end{aligned}$$

for all unobservables. In the examples considered in this work, we set $a_\tau = a_\epsilon = 300$ and $b_\tau = b_\epsilon = 6$. This was based on our previous experience with course grades.

This implies that our posterior distribution satisfies:

$$\pi(\boldsymbol{S}, \boldsymbol{\beta}, \boldsymbol{\tau}, \sigma^2_\epsilon, \sigma^2_\tau|\boldsymbol{y}) \propto f(\boldsymbol{S}|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\tau}, \sigma^2_\epsilon)\pi(\boldsymbol{\beta})\pi(\boldsymbol{\tau}|\mu_\tau, \sigma^2_\tau)\pi(\mu_\tau)\pi(\sigma^2_\epsilon)\pi(\sigma^2_\tau).$$

Using these priors insures that the full conditional distributions are known, which is the conditional distribution of a parameter given all of the other parameters in the model and the data; we define a Gibbs sampler utilizing the latent $\boldsymbol{S}$ in Section 3.1.

Indentifiability is always an issue in latent variable models. Unidentifiable models can lead to issues of convergence when attempting to fit the model (Gelfand and Sahu, 1999) and interpretation (Dawid, 1979). The existence of an intercept in Equation (1) would produce an unidentifiable model because the random effects have a non-zero mean. Following Albert and Chib (1993), the variance of the ϵ_i are identifiable because we fixed the bin boundries. If these boundries are unknown and random, we would need to fix this variance as well. For a greater discussion of identifiability in latent probit models see Huang and Bandeen-Roche (2004), Xu and Craig (2009), and Qu et al. (1996).

3. Computation

Bayesian inference revolves around the posterior distribution, the distribution of the parameters given observables. Unfortunately, computing the posterior often involves computing high dimensional integrals rarely available in a closed form. Instead, it is often easier to sample from the posterior distribution. Those samples can be used to conduct inference on parameters and to make predictions. The Gibbs sampler (Geman and Geman, 1984; Tanner and Wong, 1987) is a Markov Chain Monte Carlo (MCMC) method for obtaining a sequence of random samples from a probability distribution for which direct sampling is difficult to do. Gibbs sampling constructs a Markov Chain which has stationary distribution equal to the target posterior distribution. For an introduction to Gibbs sampling see Casella and George (1992).

3.1 Gibbs Sampler for The Probit Model

Gibbs sampling can work well when it is easy to sample from the full conditional distributions of each unknown parameter (the distribution of that parameter given all other parameters and the data). Iteratively sampling from the full conditionals builds the Markov chain. Following Tanner and Wong (1987) and Albert and Chib (1993), we use the latent grade as a parameter to facillitate the Gibbs sampler.

Recall that conditional on the observed letter grade y_i and all other parameters, S_i follows a truncated normal distribution (the letter grade tells us the range of possible values). The conjugacy of this model makes the full conditionals of each parameter known. Generally, $\boldsymbol{\beta}$, $\boldsymbol{\tau}$, are μ normal while σ_τ^2 and σ_ϵ^2 are inverse gamma.

A Gibbs Sampler is described as follows: at the tth iteration, each parameter is updated by sampling from the full conditional distributions:

$$\begin{aligned}
\boldsymbol{\beta}^{(t)} &\sim \pi(\boldsymbol{\beta}|\boldsymbol{\tau}^{(t-1)}, \mu_\tau^{(t-1)}, \sigma_\tau^{2(t-1)}, \sigma_\epsilon^{2(t-1)}, \boldsymbol{y}) \\
\boldsymbol{\tau}^{(t)} &\sim \pi(\boldsymbol{\tau}|\boldsymbol{\beta}^{(t)}, \mu_\tau^{(t-1)}, \sigma_\tau^{2(t-1)}, {\sigma_\epsilon}^{2(t-1)}, \boldsymbol{y}) \\
\mu_\tau^{(t)} &\sim \pi(\mu_\tau|\boldsymbol{\beta}^{(t)}, \boldsymbol{\tau}^{(t)}, \sigma_\tau^{2(t-1)}, \sigma_\epsilon^{2(t-1)}, \boldsymbol{y}) \\
\sigma_\tau^{2(t)} &\sim \pi(\sigma_\tau^2|\boldsymbol{\beta}^{(t)}, \boldsymbol{\tau}^{(t)}, \mu_\tau^{(t)}, \sigma_\tau^{2(t-1)}, \sigma_\epsilon^{2(t-1)}, \boldsymbol{y}) \\
\sigma_\epsilon^{2(t)} &\sim \pi(\sigma_\tau^2|\boldsymbol{\beta}^{(t)}, \boldsymbol{\tau}^{(t)}, \mu_\tau^{(t)}, \sigma_\tau^{2(t)}, \boldsymbol{y}) \\
\boldsymbol{S}^{(t)} &\sim \pi(\boldsymbol{S}|\boldsymbol{\beta}^{(t)}, \boldsymbol{\tau}^{(t)}, \mu_\tau^{(t)}, \sigma_\tau^{2(t)}, \sigma_\epsilon^{2(t)}, \boldsymbol{y})
\end{aligned}$$

The elements of $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ are updated in blocks to improve mixing. We assessed convergence by monitoring trace plots and the Gelman-Rubin diagnostic (Gelman and Rubin, 1992).

3.2 Importance Sampling for the Logit Model

Recall that in Section 2 we described how a logit model may be implemented by changing the distribution of $\boldsymbol{S}$ in the model on Equation (2) from Normal with variance σ_ϵ^2 to a logistic distribution with scale parameter σ_ϵ^2. This model is no longer conditionally conjugate, ruling out using an easy Gibbs Sampler for model fitting, but similar to O'Brien and Dunson (2004) we may use importance sampling.

Importance sampling can reduce variance and increase the efficiency of Monte Carlo algorithms for estimating integrals (Ferrenberg and Swendsen, 1988; Geweke, 1989). Instead of drawing from the target distribution, importance sampling samples from an easier

"importance" distribution. The ratio of the target distribution's density to the importance distribution's density is used as weights to estimate expectations of the target distribution.

In our example, we use the probit model (π_{probit}) as the importance distribution to estimate moments of the logit model (π_{logit}). The importance sampling estimator for the expected value of $g(\theta)$, $g(\cdot)$ is a function of any parameter θ, based on T draws from the full posterior of the probit model is :

$$\widehat{\mathrm{E}_{\pi_{logit}}(g(\theta))} \quad = \quad \frac{1}{\sum_{i=1}^{T} w_i} \sum_{t=1}^{T} g(\theta^{(t)}) w_t.$$

Where the importance weights are defined as:

$$w_t = \frac{\pi_{logit}(\boldsymbol{S}^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\tau}^{(t)}, \mu_t^{(t)}, \sigma_\epsilon^{2(t)}, \sigma_\tau^{2(t)} | \boldsymbol{y})}{\pi_{probit}(\boldsymbol{S}^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\tau}^{(t)}, \mu_t^{(t)}, \sigma_\epsilon^{2(t)}, \sigma_\tau^{2(t)} | \boldsymbol{y})}$$

and $\theta^{(t)}$ is the tth draw of that particular parameter from the probit model samples. The choice of using the probit model as the importance distribution is both because it is practical (we outlined how to obtain these draws in Section 3.1) and in our experience it works well.

4. Gain in Using the Partial Ordered Probit/Logit Models and Extensions

The latent partially ordered representation described above allows us to model all the grading information in the data. That means we do not have to throw away information by forcing the data into a logistic model or ordered probit/logit model. With our Redesigned Course evaluation project, not only can students' ordered letter grades A, B, C, D, F be used in the model, but also other partially ordered grades such as CR, NC, W, I, AUD. We can use all this information to give a more precise prediction of student performance, which is measured by a latent variable course score S_i. The predicted probability ($P(S > 73)$) of a student passing the course can be estimated based on the demographic and achievement data for a student ($\boldsymbol{X}$: SAT/ACT, Course load, Semester GPA, gender, etc.) and which instructor taught him or her ($\boldsymbol{Z}$).

With this model, some categories which do not convey much information are treated like missing data that falls within a broad range. For example, the course score for grades "I" and "AUD" can be anything between 0 and 100; we don't know these scores because the students did not finish or their performance in the course was not actually recorded. The course score of these students are predicted by the model from the data we have about these students. This is based on the work of Muthén (2004) and Muthén and Muthén (2010) who saw categorical outcomes as indicators of missingness and predicted the values with a latent variable. With this approach, the missing response data are imputed automatically by the model.

The latent grade model is easily interpreted for non-statisticians. Introducing the latent variable makes the model take the same form as a linear regression model. Non-statisticians do not have to have the knowledge about link functions and Bayesian modeling to interpret the model at high level. Some knowledge in algebra is sufficient to interpret this model.

When only binary grades are observed the partially ordered model probit/logit collapses to a standard probit/logit model. Similarily, when only ordered grades are observed, the model collapses to an ordered probit/logit model. Thus, the partially ordered framework includes both binary and ordered data as special cases.

It is easy to fit the model with Gibbs Sampling and importance sampling. Since our probit model is conditionally conjugate, the Gibbs sampler can be used to fit it, then based

on the probit model, importance sampling can be used to estimate the parameters of the logit model. In practice, estimating parameters and making predictions takes Less than 30 minutes of computer time.

5. Simulation Study

To assess properties of this model in terms of estimation, we designed a simulation study. We consider the case of $k = 2$ fixed effects and $J = 12$ different random effects. Specifically, we fixed $\boldsymbol{\beta} = (-1.45, 4.89)$, the random effect $\boldsymbol{\tau} = (81.20, 79.67, 81.80, 81.10, 75.68, 78.85, 77.14, 76.36, 76.12, 78.00, 80.91, 78.44)$, and standard deviation of error terms $\sigma_\epsilon = 6.08$. Each created observation was randomly assigned an $\boldsymbol{X}_i$ vector and assigned a group for the random effect ($\boldsymbol{Z}_i$).The first element of $\boldsymbol{X}_i$ was generated following a Bernoulli trial with probability of success 0.283. The second element of $\boldsymbol{X}_i$ was drawn from a normal distribution.

Conditional on these parameters, fixed design vector, and random design vector, each observation was generated a latent course score following Equation (1). This in turn implies a letter grade (A-F). To incorporate the partial ordering, 5% of students were assigned to earn CR and another 5% were graded NC. This process was repeated for sample sizes of $n = 500, 1000, 3000, 5000$ and $10,000$.

For each data set, we fit the model in Equation (1) with the prior distributions of Section 2 using the algorithm of Section 3. We obtained one hundred thousand post-convergence draws from the marginal distribution of each parameter and sample size. Figure 1 displays kernel density estimates of the marginal posterior distribution of β_1 and τ_7 for each sample size. The solid black vertical line represents the true value.

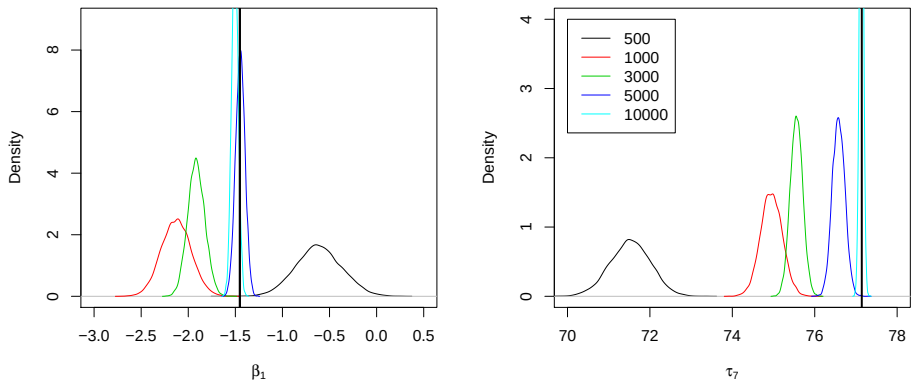

Figure 1: The left and right panel display the kernel density estimates of the marginal posterior distribution of β_1 and τ_7 respectively. Each color denotes a different sample size. The true value is represented by the solid black line.

In both panels of Figure 1, the posteriors are more variable at the smaller sample sizes. The posterior distribution from the smaller samples are also further from the true value. As the sample size increases, the posteriors concentrate more around the true value. The plots of the other parameters were so similar that we obmitted them from this work.

6. Application to our Redesigned Course Evaluation

6.1 Results

Returning to the Redesigned Course evaluation, we applied the partially ordered multinomial probit/logit models to the data. We obtained 1,000,000 post convergence draws from the posterior distribution of the partially ordred probit model using the algorithm described in Section 3, and importance sampling was used to estimate the posterior mean of all parameters in the logit model. Table 2 displays the estimated posterior mean of all $\boldsymbol{\beta}$ parameters (fixed effects) for the probit and logit models. Figure 2 displays estimated marginal posterior of the $\boldsymbol{\beta}$ parameters in the probit model.

Variable	Probit	Logit
SAT/ACT	3.732	3.073
Course load	−0.618	−0.563
Semester GPA	10.213	10.237
HS grad. years	−0.239	−0.255
High school	3.878	3.454
Gender	0.303	0.349
Asian	5.260	4.96
White	3.272	2.955
Hispanic	1.795	1.463

Table 2: The first column contains the variable name associated with each β, and column under “Probit" and “Logit” are the estimated posterior means of the elements of $\boldsymbol{\beta}$ for each independent variable, respectively.

From the table, we can see that the parameter estimates from the two models are similar. The coefficients for SAT/ACT and semester GPA are positive and large, which means students who did well on the SAT and ACT or have a higher GPA in the semester tend to do better in the Course. Course load has a negative coefficient, indicating that students who took more courses tended to do worse in the Course. High school graduates did better than GED certificate holders. Asian students did a little bit better than Whites, Whites did better than Hispanics, and Hispanics did better in the Course than other races. Gender and high school graduation years did not appear to be significant predictors.

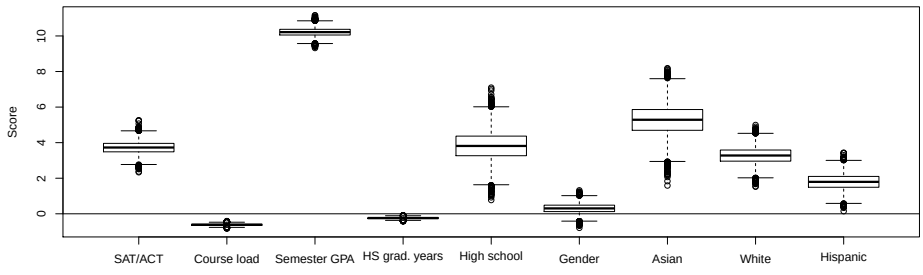

Figure 2: Boxplots of $\boldsymbol{\beta}$ estimated using draws from the posterior distribution.

Figure 3 displays boxplots of the marginal distribution of the instructor random effects under the probit model. The random effects were centered about zero for easier comparison.

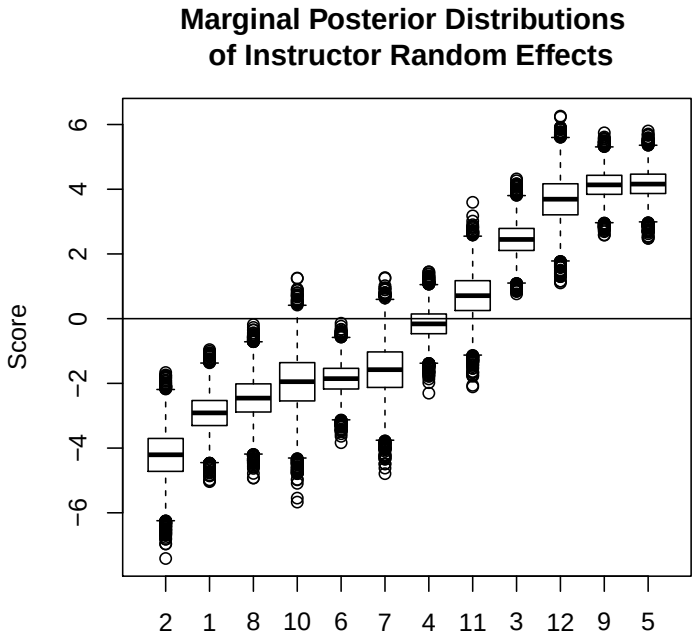

Figure 3: Boxplots of instructor random effects (τ). The draws were centered by subtracting the overall mean (of all instructor effects). Instructor 12 is the Redesigned Course and the numbers 1-11 represent different instructors who taught the traditional lecture sections.

Recall that we were primarily interested in how students in the Redesigned Course (the experimental section) performed relative to other students. Looking at the box plots, we see that the model estimates that the grade distribution in the Redesigned Course (listed as Instructor 12) is similar to the instructors with the highest observed grade distributions (instructors 9 and 5) and higher than most of the other instructors.

6.2 Prediction with The Paritially Ordered Probit/Logit Model

An advantage of fitting models through Markov Chain Monte Carlo approaches is that we can easily make predictions about the course score for individuals. Consider a student who is a white, male, high school graduate with average SAT/ACT score (the ACT Math score 19 or SAT Math score 463), average Semester GPA (2.97), average number of years since high school graduation (2.6) and who was in the Redesigned Course. Using our fitted probit model, his average predicted score is 80.3 and his probability of passing the Course is 0.75. Figure 4 displays the posterior predicted grade distribution for this student under both the probit and logit model.

If this student was instead taught by instructor 2, his average posterior predicted score under the probit model would change to 72.2 while his probability of passing the course drops to 0.47. Figure 5 displays the posterior predicted grade distribution for this student under both the probit and logit model.

If the model is validated by future grade distributions that include the final grades for students who received an Incomplete, this suggests that the ability to quickly make predictions for students allows us to quantify how much better a section would perform if the format was changed to the Redesigned Course.

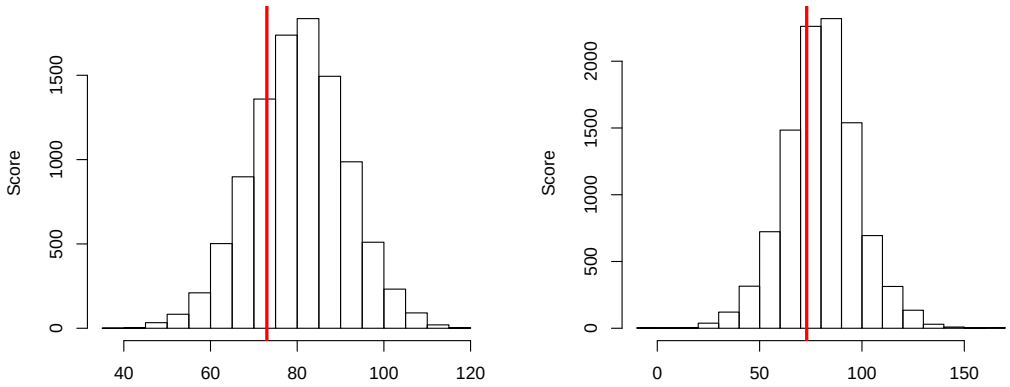

Figure 4: Estimated posterior predictive distribution of scores for an example student from the Redesigned Course section under the probit (left figure) and logit (right figure) model. The vertical red line represents the passing score.

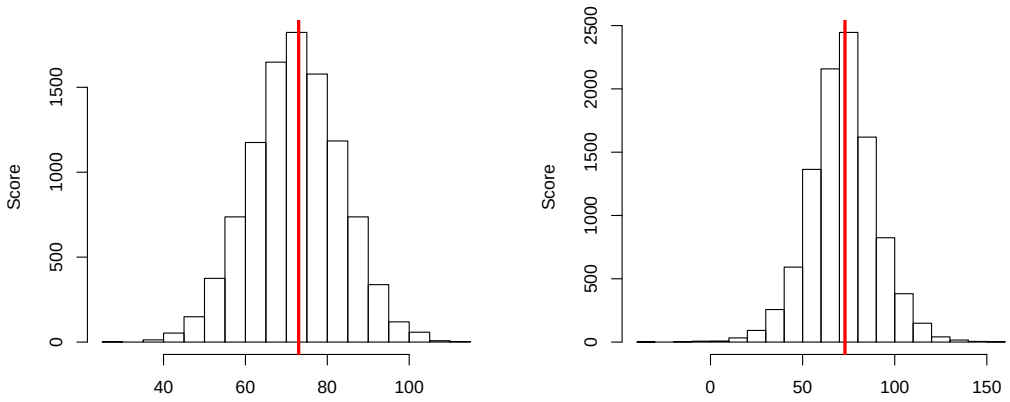

Figure 5: Estimated posterior predictive distribution of scores for an example student from instructor 2's section under the probit (left figure) and logit (right figure) model. The vertical red line represents the passing score.

6.3 Comparison to Alternative Analyses

The main advantage of this model is that we can use all of the data while an ordered or standard probit model cannot. In our example, the standard probit model ignores the magnitude of the letter grade by converting them to passing or failing grades and completely ignores those who recieve a W, I or AUD. The standard ordered probit model similary ignores grades of CR, NC, W, I, AUD. The ordered probit uses the least amount of observations but has more detailed information than the standard probit model. The partially ordered probit model uses all of the observations with all of the detail.

To quantify the impact of this loss of information, we utilized five-fold cross validation with 2 measures of model fit: a mean squared error (MSE) and the area under the ROC

curve (AUC). These measures are based on only predicting the probability of passing the course. We chose this because the University is most interested in predicting pass rates. For the ordered probit and standard probit, we used the Bayesian formulation from Albert and Chib (1993) with the same random effects and predictors as the partially ordered model.

The data was randomly placed into 5 folds of equal size. For each of the five folds, the data not in that fold was used to fit each of the three probit models. Predictions of the passing rate were made for each student in that fold. From those predictions, the MSE and AUC were calculated with the actual pass/fail results for those students. Students who received a W, I, or AUD were excluded at this stage. Table 3 presents the average (over the five folds) MSE and AUC for all three models.

Model	MSE	AUC
Partially Ordered Probit	0.310	0.654
Ordered Probit	0.327	0.599
Probit	0.348	0.642

Table 3: The first column lists the models used. The second column gives an estimated MSE using five-fold cross validation. The third column gives an estimated AUC using five-fold cross validation.

We see from the table that, in terms of MSE and AUC, the partially ordered model performs the best. In other words, the information gained from using all grading information for students does translate into a better fitting model. The ordered model performs worse than the standard probit for this data in terms of AUC but better in terms of MSE. We attribute this to it being unable to use the 158 students who received a CR or NC (no letter grade). The extra detail of the ordered model could not compensate, in terms of model fit, for the loss of these students. If there were no CR/NC students, we suspect that the ordered model would perform better than the standard probit model.

7. Discussion

Student success is often measured using letter grades, which are sometimes totally ordered but often have only a partial ordering. The Bayesian Partially Ordered Probit/Logit Models presented in this work are a way of retaining all of this ordering information. These models are generalizations of the standard models in the sense that totally ordered or binary response variables, which are required for the standard models, are special cases of partially ordered response variables. We used a Bayesian approach because prior knowledge about the parameters can be incorporated into the model and the Gibbs sampler can easily to fit the probit model and a logit model can be used with importance sampling. Modifications to the model in Equation (1) can easily be made. Say we observe student performance in courses for which the Course in question is a prerequisite: autoregressive error terms could be utilized in this case to model students' performance in all future courses. Additionally, the model we present is understandable to non-statisticians.

The Partially Ordered Probit model out performs totally ordered and binary probit models in terms of cross-validated mean squared error and area under the ROC curve. This is not surprising since it includes more information (the partially ordered subjects) in the same general model structure. In a sense, the model automatically performs an imputation on the partially ordered subjects. The gain we see is thus related to the gain in modeling the "missing" data.

There are, however, limitations to the study and clear areas of needed future work. The model has the potential to help fill in the missing data represented by the large number

of Incompletes; however, the method we used assumed that Incompletes were randomly assigned to students, when this may not be the case. While it is true that if a student gets an Incomplete in a standard section where such grades are reserved for extreme situations like an illness or family emergency, it may not be the case in a self-paced course that builds such grades into its structure. In other words, it is reasonable to assume that the Incomplete grades are randomly assigned in a standard section, but they may reflect something about a student's study skills, for example, in the Redesigned Course. Settling this issue will require examining the actual grades that students eventually receive in the course, which is data that was not available at the time of the study.

We developed this model specifically for the Redesigned Course evaluation discussed in Section 1, but it is easily extended to other categorical data problems with a partially ordered response variable. For example, criminal sentences are partially ordered: one year in prison is clearly less than two years. But one year of prison with 3 years probation is not necessarily less than 16 months in jail. A data set containing sentences, crimes, and demographic characteristics of the criminal could be analyzed with this model. Our model can be applied to this and other kinds of partially ordered categorical data.

References

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422):669–679.

Casella, G. and George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174.

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

Ferrenberg, A. M. and Swendsen, R. H. (1988). New Monte Carlo technique for studying phase transitions. *Physical review letters*, 61(23):2635.

Fullerton, A. S. (2009). A conceptual framework for ordered logistic regression models. *Sociological methods & research*, 38(2):306–347.

Gelfand, A. E. and Sahu, S. K. (1999). Identifiability, improper priors, and Gibbs sampling for generalized linear models. *Journal of the American Statistical Association*, 94(445):247–253.

Gelman, A. and Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 6:721–741.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, 57(6):1317–1339.

Gilmour, A., Anderson, R., and Rae, A. (1985). The analysis of binomial data by a generalized linear mixed model. *Biometrika*, 72(3):593–599.

Hedeker, D. (2008). Multilevel models for ordinal and nominal variables. In *Handbook of multilevel analysis*, pages 237–274. Springer.

Huang, G.-H. and Bandeen-Roche, K. (2004). Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika*, 69(1):5–32.

Kaplan, D. and Elliott, P. R. (1997). A model-based approach to validating education indicators using multilevel structural equation modeling. *Journal of Educational and Behavioral Statistics*, 22(3):323–347.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2):109–142.

McKelvey, R. D. and Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of mathematical sociology*, 4(1):103–120.

Muthén, B. (2004). Latent variable analysis. *The Sage handbook of quantitative methodology for the social sciences. Thousand Oaks, CA: Sage Publications*, pages 345–68.

Muthén, L. K. and Muthén, B. (2010). *Mplus*. Muthén & Muthén Los Angeles, CA.

O'Brien, S. M. and Dunson, D. B. (2004). Bayesian multivariate logistic regression. *Biometrics*, 60(3):739–746.

Qu, Y., Tan, M., and Kutner, M. H. (1996). Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics*, 52(3):797–810.

Raudenbush, S. (1988). Educational applications of hierarchical linear models: A review. *Journal of Educational and Behavioral Statistics*, 13(2):85–116.

Schaafsma, J. and Osoba, D. (1994). The Karnofsky performance status scale re-examined: a cross-validation with the eortc-c30. *Quality of Life Research*, 3(6):413–424.

Stiratelli, R., Laird, N., and Ware, J. H. (1984). Random-effects models for serial observations with binary response. *Biometrics*, 40(4):961–971.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540.

Winship, C. and Mare, R. D. (1984). Regression models with ordinal variables. *American Sociological Review*, 49(4):512–525.

Wong, G. Y. and Mason, W. M. (1985). The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association*, 80(391):513–524.

Xu, H. and Craig, B. A. (2009). A probit latent class model with general correlation structures for evaluating accuracy of diagnostic tests. *Biometrics*, 65(4):1145–1155.