4-2013

Are You Paying Attention? Understanding Student Attention Allocation using Bayesian Models

Elizabeth C. Lorenzi
Carnegie Mellon University

Are You Paying Attention? Understanding Student Attention Allocation using Bayesian Models

Elizabeth C. Lorenzi

April 29, 2013

Abstract

A mixed effects multinomial logistic model is useful in understanding a response variable with more than two outcomes and its relationship with covariates for nested data sets. Because of the nested structure of the data, using random intercepts is needed to adjust for the dependency within subgroups. This paper addresses this method while analyzing a psychological study performed by the Carnegie Mellon Psychology Department. The data are collected from local elementary schools to better understand the environmental effects that cause off task behavior in the classroom. Logistic multinomial models with Bayesian analysis are used to better understand what behavior and activity is responsible for making students go off task in the classroom.

1 Introduction

Bayesian analysis is a realm of statistics that has become more popular for use on social sciences research. It allows for the use of known prior information to help determine the outcome of relationships between the response variables and the covariates. Alternatively, Bayesian analysis can use diffuse priors to analyze data and can be more effective than Frequentist approaches when missing data exist in the data set or small sample sizes come into play. By using the likelihood from the data and prior distributions, we calculate a posterior distribution of our end result using the computational method, Markov Chain Monte Carlo. Coding a model not available in standard software (e.g. the hierarchical multinomial) is often easier with WinBUGS than in R for a frequentist approach because you are not required to derive the formulas for the standard errors.

In this paper, we analyze our data using Bayesian analysis for multinomial logistic regression with mixed effects. Though research exists on calculating multinomial, Bayesian models, our paper explains the methodology of using these models using the program WinBUGS and the R package, Rube, written by Dr. Howard Seltman. We discuss a process of taking a nested data set with more than two response categories and describe the modeling, diagnostics, and model-checking to best find the appropriate model to analyze the data.

The starting point for this project is from a classroom study that aims to understand what environmental effects cause students to be off task in the classroom. The study is performed by Dr. Anna Fisher from the Carnegie Mellon University Psychology Department in five different schools in the Pittsburgh area. It focuses on students between the grades, kindergarten to first grade. The data are collected by observers in the classroom, denoting the time of the observation, the behavior of the child during the observation, and the activity being performed. Using these pieces of information we hope to see how time and activity affect a child's attention in the classroom and specifically what way the student becomes off task. The data are collected in sessions, where each student could have between 1 and 30 observations.

Because the data are collected over time, we believe time will have an effect on whether a student is on or off task. Our hypothesis is that over time a student will be more likely to go off task and regain attention during a change in activity or the start of class. Because of this we research potential time decays on the log odd scale including linear, exponential, and quadratic.

It is important to model in steps and find the best approach through stages. This paper does that first beginning with simulating data and finding an appropriate way to model time decay, next we move to the real data and model in binomial terms, lastly we write the model to predict the multinomial response. Through these steps we display diagnostics, model-checking, and model fitting, with the end result leading us to understand what aspects of the classroom affect a child's attention allocation.

2 Literary Review

We begin by analyzing work done by others in the topic of Bayesian hierarchical binary and multinomial logistic models as well as different research on parameter priors. To understand how to approach the goal of analyzing nested, multinomial data with Bayesian models it is important to gather different opinions on priors and model-fitting. Throughout our analysis we use the advice from Andrew Gelman in using noninformative or diffuse priors that result in posterior distributions that describe the likelihood more strongly than the prior information [2]. Additionally, Andrew

Gelmen wrote *Prior Distributions for Variance Parameters in Hierarchical Models*, where he suggests appropriate non-informative prior distribution for scale parameters in hierarchical models [1]. Specifically he focuses on the variance parameter priors and suggests a uniform prior on the hierarchical standard deviation and a half-t-family when the number of groups is small, and warns against the inverse gamma distribution because of its tendency of miscalibration and its result of an improper posterior distribution.

In the *Hidden Dangers of Specifying Noninformative Priors*, Seaman III, Seaman Jr. and Stamey discuss the potential downfalls of using diffuse priors and possible solutions to avoid them [5]. Often noninformative priors have an unintended influence on the posterior of the function such as shrinkage and surrogacy. They propose solving these problems using simulation to check that the priors are not influencing the posterior away from the values simulated, performing a prior-posterior sensitivity analysis using plots of the posterior and prior, and checking that the posterior results are similar to the maximum likelihood results (unless prior beliefs are strong relative to the amount of data collected). These solutions become very influential during our modeling using diffuse priors.

Donald Hedeker's *A Mixed-Effects Multinomial Logistic Regression Model* describes the modeling of multinomial logistic regression with mixed effects, which was helpful as we implement this type of model using Bayesian inference [4]. The article dives into how to estimate the random effects for both ordinal and nominal response data, providing the probability equations for each group as exponentiated log odds of a given group versus the baseline group divided by 1 plus the sum of all of these exponentiated log odds for all groups. He illustrates his methods using analysis of a psychiatric data set about homeless adults with mental illnesses, grouping the data by their living arrangement. These three articles in particular were strong influences on my analysis providing important information on non-informative priors for variance parameter priors, methods on how to check for potential problems seen when using such non-informative priors, and information on the best way to calculate multinomial models. Our paper builds from all three sources to illustrate the methodology used to analyze nested multinomial data using Bayesian multinomial mixed effects models.

3 Classroom Study Data Exploratory Analysis

The data were collected through a study performed by Dr. Anna Fisher from Carnegie Mellon's Psychology Department in order to study what proportion of off-task behavior is attributed to environmental distractors in elementary and kindergarten classrooms. It is known that as students mature, they are less distracted in the classroom. Therefore the study focuses on students in the early stage of education, grades K-4.

Trained coders are responsible for observing the students, with training to define on task and off task behavior in different learning environments doing different actions. The observations are carried out by 1-2 coders at a time, sitting in an area where all of the students faces can be seen. Each observation is taken within 20 seconds, coding whether the student was on task or off task doing different behaviors.

We will move forward by looking at the data gathered from the first three waves of the psychology study performed by Doctor Fisher's team, where they visited five different high schools in the Pittsburgh area and observed students in different classrooms. Within these five schools there are 22 classrooms with 745 students. Coders visited these classrooms and noted the students attention allocation during 40 different sessions where they recorded students and their behavior during different classroom activities. In Figure 1, we see a set of boxplots to show the number of observations in each session. This plot illustrates the high variability of number of observations per session, ranging from less than 5 to 25.

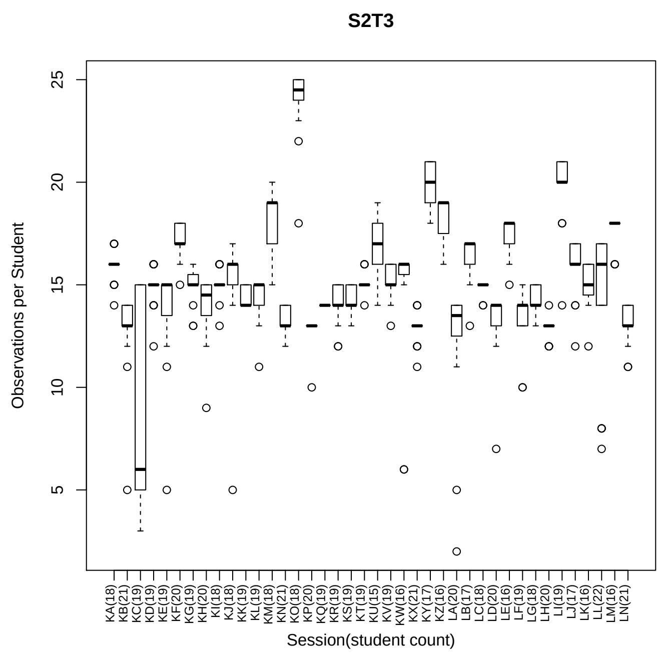


Figure 1: *Boxplots of the number of observations per session.*

Other interesting features to note of the data are the time figures they recorded. We have each observation's exact time stamp including day, hour, minute, and seconds. With these we are able to see whether over time, students attention decreases, a hypothesis we plan to continue to study in the data using the decay models discussed in later sections. It is interesting to note that 86.5% of the observations took place in the morning versus the 13.5% in the afternoon. We adjust the time variable for our analysis to measure the time of the observation after the new session begins and when an activity changes during the session. Our goal is to measure the attention allocation of students over time and see whether longer time during one activity will decrease a student's attention. By restarting the time measurement with each change in activity, the model assumes that a student returns to his/her baseline at the start of the next activity.

During each session the coders noted the activity occurring in the classroom, which are the categories: (1) Individual, (2) Other, (3) Small group individual, (4) Small group teach, (5) Whole carpet, and (6) Whole desks. During these activities, the coders observed the students behaviors, placing them in categories: (1) Self-distraction; (2) Peer-distraction; (3) Environmental distraction, (4) Sleeping, (5) Moving around the classroom, and (6) Other. Within these categories, it is interesting to note that students were on task 67.3% of the time, with the second largest category being peer with a percentage of 14.4%. This corresponds to an odds of peer behavior versus the on-task baseline of $14.4/67.3 = 0.21$ or a log odds of -1.54. In a model without covariates, this is the value we would expect for the peer intercept in a multinomial model.

In Figure 2 we see a principle component analysis biplot to show the activities that have the most influence on the data. We again see that the on task dimension is the strongest, meaning that this first principal axis maximizes the variance when the data are projected onto the line. The second longest axis is peer, denoting that peer maximizes the remaining variance compared to the other arrows shown in the plot. Because we see the other five categories are all pointed in the same direction, this means there is no strong difference between them and together maximizes the variance in that direction. This leads us to make an "other" behavior category that now encompasses: (1) Self-distraction; (3) Environmental distraction, (4) Sleeping, (5) Moving around the classroom, and (6) Other.

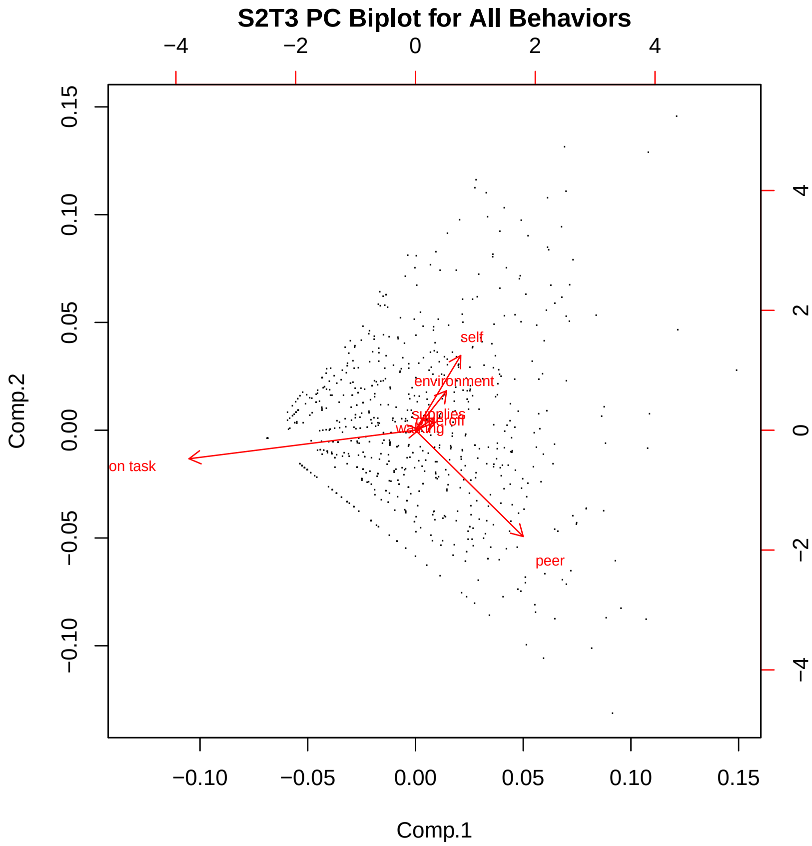
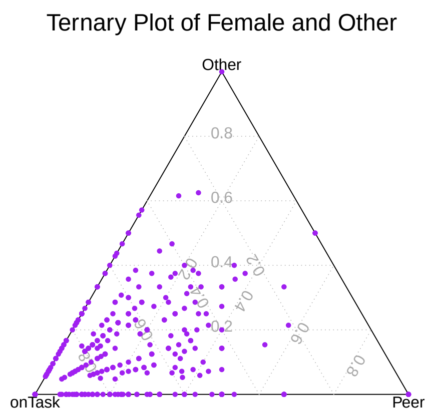
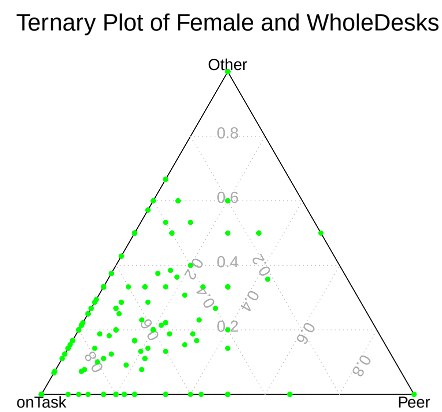
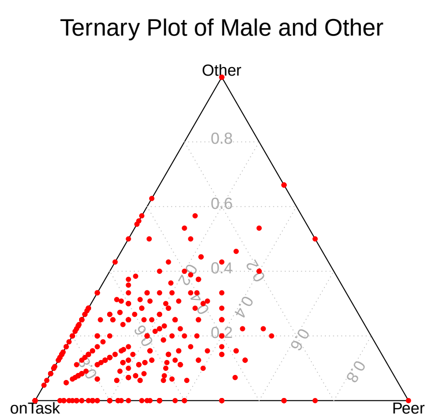


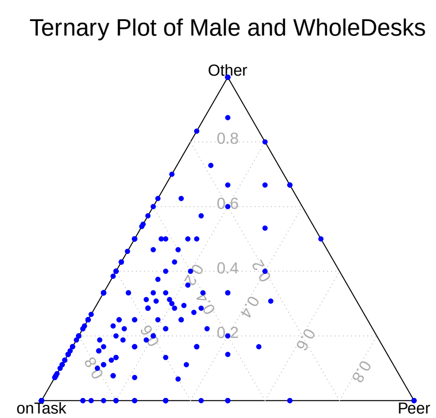
Figure 2: *Biplot of principle component analysis on behavior.*

Additionally, using ternary plots we are able to see how gender and activity can affect the response variable. We see in female and other and male and other that the plots look very similar with the majority of the points hovering around the on task corner signifying the greater probability of being on task for the other activity category. When the activity is whole desks, we see that the points spread out more and move closer to the off task with other and off task with peer corners, signifying that whole desks may be an activity that causes students to move off task. For the plot of males and whole desks, we see that the points are spread even more across the triangle, but hovering closer to the edge between other and on task. This signifies that there is little off task with peer behavior and students that are male and are doing the activity whole desks are more likely to either be off task with other or on task. These ternary plots give us a good starting point to understand how some of the activities affect behavior and whether there is a gender effect that affects students attention as well. Because we saw little difference in the plots between male and female, it leads us to not model with an interaction between gender and activity.









4 Data Simulation

4.1 Investigation: Decay of On-Task Behavior in a Hierarchical Logistic Regression Model

The first step of the research is to simulate data mirroring the psychology study of measuring on task behaviors in the classroom to investigate the effectiveness of various models and then apply them to the real classroom data. The study is to track the students' on task and off task behavior (collapsed over all of the off-task categories) over several observations in a specific time frame. We have students from various grades K-5 from different school districts in the area of Pittsburgh. To begin the analysis, we simulate data matching these characteristics to find a method that will match the features of the real data and then model the data to predict the on task behavior of students.

The response variable signifies a student on task or off task measured 10 times during an observation period, simulated by 10 measurements of 0 or 1 for each student. Because we believe that there is an intrinsic decay of attention over a time period within students, we compute the response variable using a linear decay over time on the log odds scale. To test other hypotheses that the attention decay can also be modeled through an exponential decay or a quadratic decay, we create two other data sets that we will use with separate models testing the exponential and quadratic decay of attention. Additionally, we simulate covariates such as gender, age (6-12), and economic status (high, medium or low).

The equation in (1) shows the logistic formula using the covariates female, age, economic status, and time. We model time in this case as a linear decay all combining to determine the log odds of success.

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_{fem}\text{fem} + \beta_{age}\text{age} + \beta_{hi.econ}\text{hi.econ} + \beta_{med.econ}\text{med.econ} \quad (1)$$
$$-\beta_{time}time + true.RI \sim Norm(0, \tau^2)$$

To simulate, we provided values for these covariates so to calculate the log odds of y. The coefficients for these covariates are as follows: $\beta_0 = 1$, $\beta_{time} = 0.25$, $\beta_{fem} = 0.5$, $\beta_{age} = 0.1$, $\beta_{hi.econ} = 0.5$, $\beta_{med.econ} = 0.5$. We draw a random intercept in the log odds equation, so that for each student we have an individual intercept. Next, we convert the log odds into probabilities based on the equation (1), allowing us to randomly draw from the binomial distribution 10 times for each student based on these probabilities of success.

The resulting data are composed of 200 students with 10 measurements per student of on task behavior, the three covariates described, the index number of the student, and a column for the number of observations for student.

4.2 Modeling: Linear Decay of On Task Behavior on the Log Odd Scale

We model the data using random intercept, binomial, Bayesian models through the use of Rube and WinBUGS. Before moving into the classroom study data, we want to make sure we write models that analyze the data accurately. This can be assessed by checking the accuracy of the coefficients in our models compared to the coefficients we used to simulate. We begin by writing a linear decay on the log-odd scale model to match the simulated data that reflects the linear decay. This is done using a binomial model to predict y (on task- 1, off task- 0) based on the variables, gender, age, medium economic status, and high economic status. We assign priors to the covariates gender, age, medium economic status and high economic status (low economic status as the baseline), using a normal distribution with mean 0 and standard deviation 1000. We chose a weakly informative prior for these to assure the MCMC method will converge and that the posterior parameter distributions will not be influenced by the prior mean to any meaningful degree. This should result in a distribution that reflects our data very similarly. We model y as a binomial with probability derived from the log odds function with a random intercept for each student. The random intercept is modeled for each student with a normal distribution with mean zero and unknown precision, which is modeled as a uniform between 0 and 10 on the standard deviation scale. Y is modeled for each observation using a binomial distribution, resulting in 10 individual predictions for each student. The binomial distribution is based on the probabilities of success from the linear log odds function, with a common slope and a separate random intercept for each student. Below in equation (2), the model is described in hierarchical Bayes formatting as well as WinBUGS code.

$$
\begin{aligned}
y_i &\sim \text{Binomial}(\text{inverse.logit}(\alpha_{0ji} + \beta_{fem}\text{fem}_i + \beta_{age}\text{age}_i + \beta_{hi.econ}\text{hi.econ}_i \\
&\quad +\beta_{med.econ}\text{med.econ}_i - \beta_{time}\text{time}_i)) \qquad (2)\\
\alpha_{0ji} &\sim \text{Normal}(\beta_0, \tau^2)\\
\beta_{femi} &\sim \text{Normal}(0, 1000)\\
\beta_{agei} &\sim \text{Normal}(0, 1000)\\
\beta_{med.econi} &\sim \text{Normal}(0, 1000)\\
\beta_{hi.econi} &\sim \text{Normal}(0, 1000)\\
\beta_t &\sim \text{Normal}(0, 1000)
\end{aligned}
$$

Rube Code:

```
decay.m = "model {
  for (i in 1:N) {
    y[i] ~ dbin(p[i], n[i])
```

```
    logit(p[i]) <- LC(b, COVA, , i) - Bt*time[i] + ri[id[i]]
  }
    for (j in 1:NS) {
      ri[j] ~ dnorm(0, ri.prec)
    }
    ri.prec <- pow(ri.sd,-2)
    ri.sd ~ dunif(0,RI.SD.MAX=10)

  FOR(b, COVA, , ?~dnorm(0, COV.PREC=1e-6))
  Bt~dnorm(0, BT.PREC=1e-6)
}"
```

The COVA in the rube code are any subject-level covariates that may affect individual intercepts, with the LC being a rube command that fits all of the variables in COVA as coefficients in the model. Further in the code, the use of all capital FOR is a rube command that says for all the variables in COVA, to set a prior normal distribution with mean 0 and precision $1e^{-6}$. The coefficients initial starting values are drawn randomly from various normal distributions with mean 0 and wide standard deviations to set starting points for the MCMC chains. After checking that the initial values and prior distributions are as intended, we run the model saving the posterior parameter estimates β_0, β_{fem}, β_{age}, $\beta_{hi.econ}$, $\beta_{med.econ}$, β_t, and the random intercepts. Using diagnostic plots (Rube p3 function), we see that the three MCMC chains do not converge until the 1000th iteration, leading us to discard the first 1000 iterations to get closer to the period of convergence of the MCMC chains, shown without the burn in in Figure 3. Next, using the autocorrelation plot, we detected low lag correlation within the estimates, leading us to thin to every tenth iteration. We decide on the 3000 total iterations because with the first burned and then thinned by ten, we are left with a total of 200 iterations, which gives a negligible MCMC error compared to the posterior standard deviations These iterations lead to convergence signified by the Gelman R-hat statistics near 1, low values for MCMC error, and evidence of good mixing in the trace plots of posterior values versus iteration number [2]. Below is a plot of β_0's result, representing the mean intercept of the model. The plot shows the MCMC chains convergence over the iterations, the autocorrelation plot, and the density of the three MCMC runs for the three different starting values. We see that the MCMC curves appear to converge relatively well around a mean of 1.2 for the intercept with a 95% posterior interval of approximately 1.0 to 1.4. The autocorrelation plot shows some remaining correlation among the points, so further thinning could be suggested. Lastly, the third plot shows that the centers of the MCMC chains are relatively close to one another at 1.2.

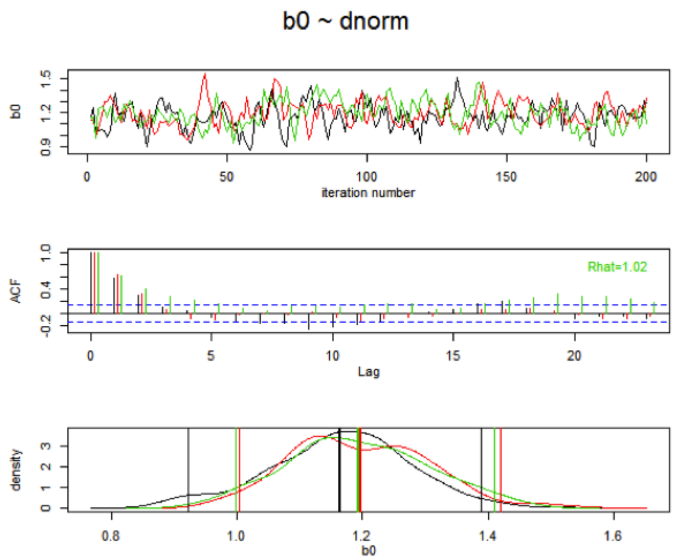


Figure 3: *Rube output for intercept estimate*

To summarize the random intercepts of the model, we display Figure 4 that includes 20 out of the 200 random intercepts per student. The plot shows that our random intercepts all converge with a Gelman R-hat value of 1. Figure 5 shows a brief summary of the remaining covariates: gender, age, and economic status. This plot shows that again the three chains converge finding Rhat values of approximately 1 and seeing little autocorrelation. We also look at the results from the time coefficient, β_t in Figure 6. This shows that this coefficient reaches convergence and has no correlation across observations after thinning every 10th observation. We see that the time posterior mean is 0.30, with a posterior credible interval that does not include zero. This means that our model found a significant time effect in the data.

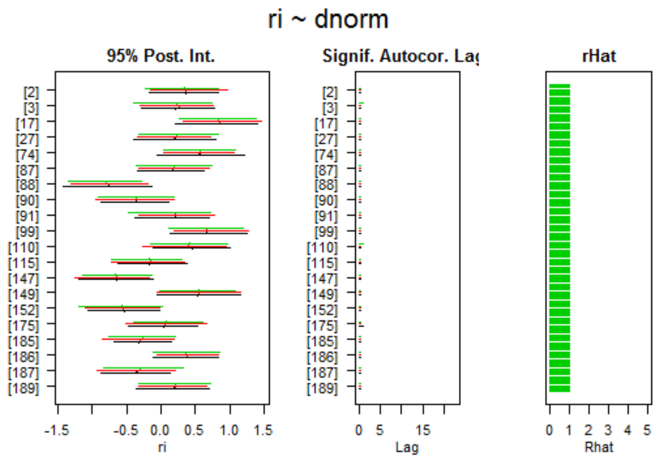


Figure 4: *Rube output for random intercept estimate per student*

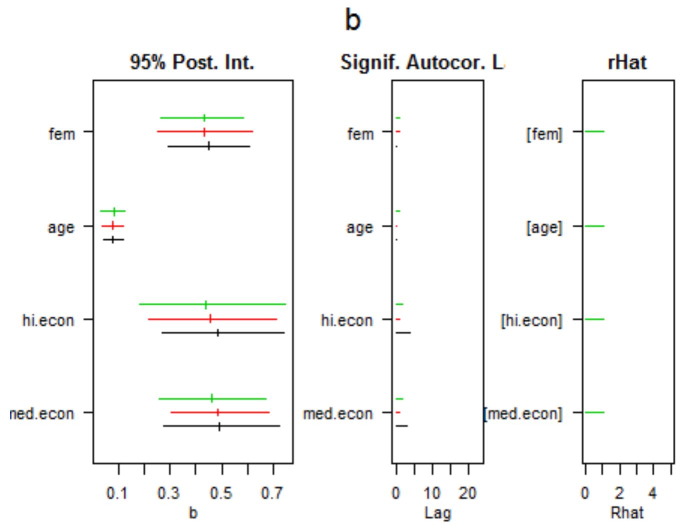


Figure 5: *Rube output for covariate estimates: Female, Age, Hi.Econ, Med.Econ*

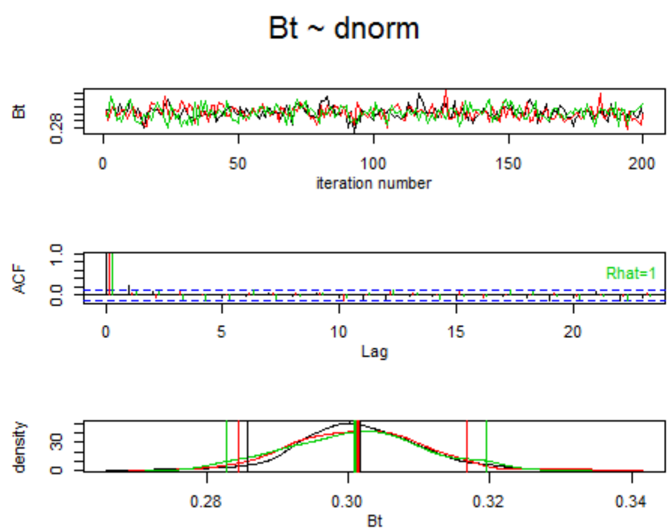


Figure 6: *Rube output for time estimate*

Table 1 reaffirms that our model fit the data relatively well, matching the initial coefficient values we set for the data simulation very closely, $\beta_0 = 1$, $\beta_{time} = 0.25$, $\beta_{fem} = 0.5$, $\beta_{age} = 0.1$, $\beta_{hi.econ} = 0.5$, $\beta_{med.econ} = 0.5$.

Table 1: Saved parameters: posterior distributions

	Mean	SD	MCMC error	Credible Interval	Simulated Values
β_0	1.12	0.11	0.00	$(0.89, 1.33)$	1
β_{fem}	0.43	0.09	0.9	$(0.25, 0.61)$	0.5
β_{age}	0.08	0.02	0.2	$(0.04, 0.12)$	0.1
$\beta_{hi.econ}$	0.46	0.13	0.13	$(0.21, 0.74)$	0.5
$\beta_{med.econ}$	0.48	0.01	0.11	$(0.26, 0.59)$	0.5
β_t	0.30	0.01	0.01	$(0.28, 0.32)$	0.25
$\beta_{ri.sd}$	0.51	0.01	0.80	$(0.28, 0.32)$	0.8
$\beta_{ri[1]}$	-0.52	0.28	0.01	$(-1.09, 0.04)$	
$\beta_{ri[10]}$	-0.16	0.28	0.10	$(-0.73, 0.35)$	

To better analyze the strength of the data relative to the prior distribution, we can compare the prior and posterior distributions of the random intercepts [5]. This can be applied to all the covariates' posterior distribution, but we will use the random intercepts because it is most interesting because of the controversy over choosing prior distributions for variance parameters [1]. In Figure 7, we show the comparison between the prior and the posterior distribution for several students from our model. We can see a significant difference between the prior distribution and the posteriors, where the posterior of the three students are much more informative than the prior all centered between about -2 and 2. The blue line represents the student with the largest mean random intercept centered around 1.3. This suggests that our prior distribution is an appropriate non-informative starting point and the chosen prior does not overly influence our model. To check the sensitivity to the prior, we change the prior of the β_t distribution to a normal distribution with mean zero and precision 1e-6 versus the previous distribution of mean zero and precision of $1e^{-4}$, resulting in a less informative prior. We see in Figure 8, that weakening the prior of β_t to $1e^{-6}$ produces posterior distributions very similar to the original prior, meaning that by weakening the prior, the distribution of the posterior parameters are not affected. This leads us to conclude that the precision of our original prior is appropriately non-informative, as we intended for this data analysis.

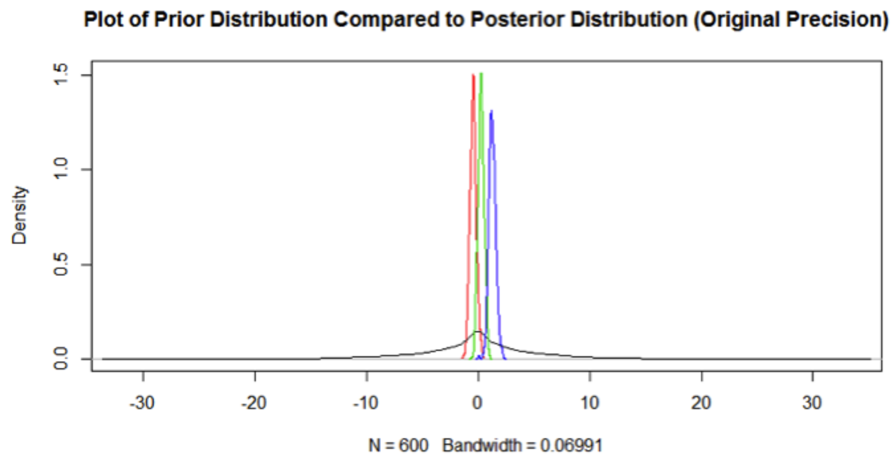


Figure 7: *Original Prior/Posterior Plot for some specific per-subject intercepts (black=prior, colors= random intercepts*

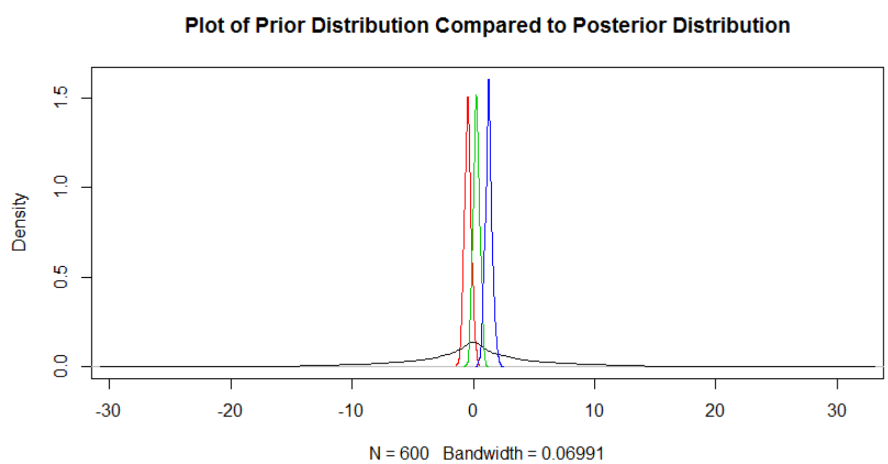


Figure 8: *Weaker Prior/Posterior Plot for some specific per-subject intercepts (black=prior, colors= random intercepts*

We can take a closer look at the random intercepts in Figure 9. We plot the random intercepts created in the simulation against the random intercepts fitted from the model. We see that the random intercepts when compared to the linear model of estimated values on true is approximately scattered evenly around the line. The slope of the line appears to be less than 1, meaning that the estimated intercepts are slightly smaller on average compared to the true intercepts. This reflects the shrinkage effect towards the mean [2].

Figure 9: *Random intercepts versus true intercepts*

As part of our sensitivity analysis for the prior's variance parameter for the covariates (COV.PREC in rube code), we compare the model using the two different precision values, the original sd= $1e^{-6}$ and a larger sd= $1e^{-4}$. In Figure 10, we see the results of the posterior distributions for the intercept using the two separate precision values. Both predictions reach convergence, shown by the Rhat values on the autocorrelation plot. Each variable shows a similar pattern in the density distribution plot, displaying both models ending in similar distributions. This again reaffirms that weakening the prior distribution is not necessary and the original prior is appropriately weakly-informative, allowing the likelihood to be more influential in the posterior distribution. We also look through these plots for the other covariates and the error term. We see a very similar pattern for each variable.

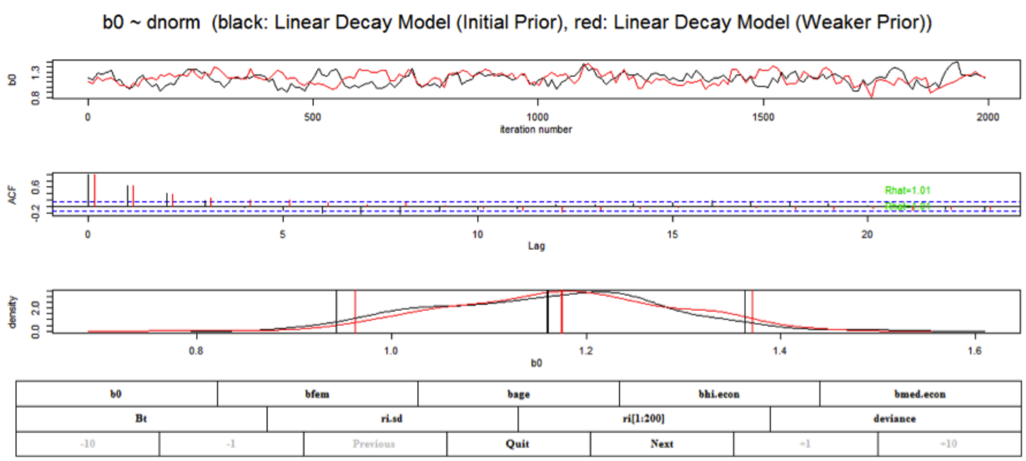


Figure 10: *Comparison between initial prior (precision 1e-4) and weaker prior (precision 1e-6)*

We will move forward working with the original model with the prior Normal(0, precision=$1e^{-6}$). Next, to check further diagnostics of our model we look to residuals. Logistic regression is a non-linear model that does not hold the same assumptions of that of a linear model. However, we can better understand the residuals by binning them as suggested by Gelman and Hill [2]. Additionally, to understand the patterns of the per student observations we plot each student's residuals. The first plot shows the residuals versus time values of the model, by looking at the first 50 students to check the residuals per student. Overall, Figure 11 shows a random scatter and minimal patterns in the per student plots, signifying that the model is fitting the data appropriately overtime and the time decay seems to be appropriately modeled. Additionally, it is informative to examine all the residuals versus time, which looking at the plots in Figure 12 we see that both the normal residuals and the binned residuals look approximately random around 0. This allows us to confirm that the model is fitting time appropriately, not seeing any certain pattern over time.

The simulation results and diagnostics prove that the model is effective at predicting the posterior distribution.

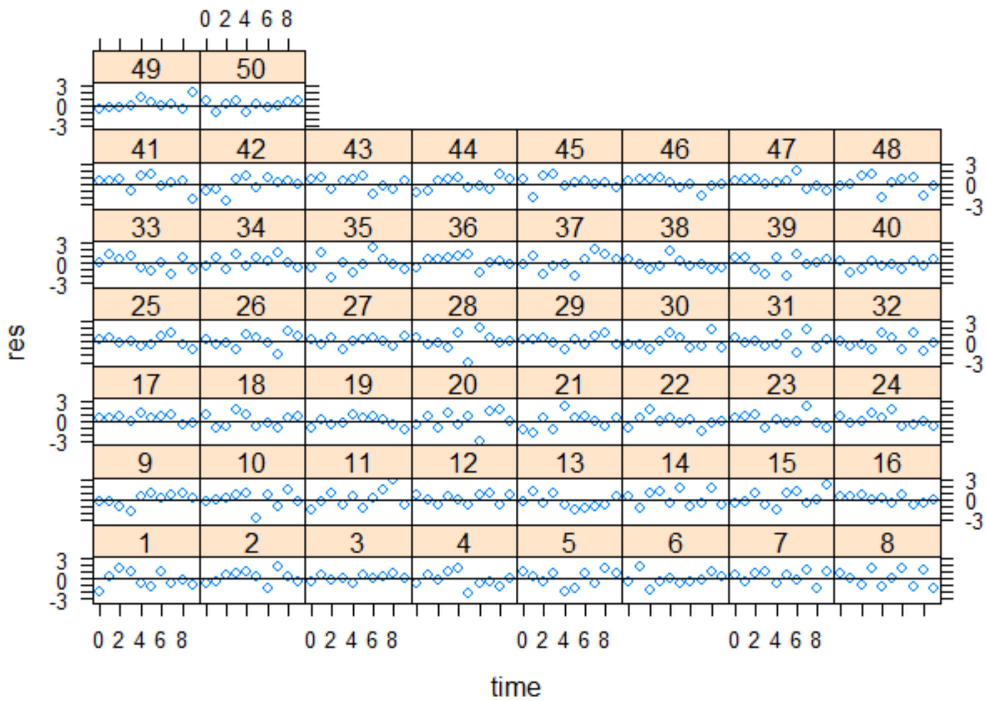


Figure 11: *Residuals versus time per student (first 50 students)*

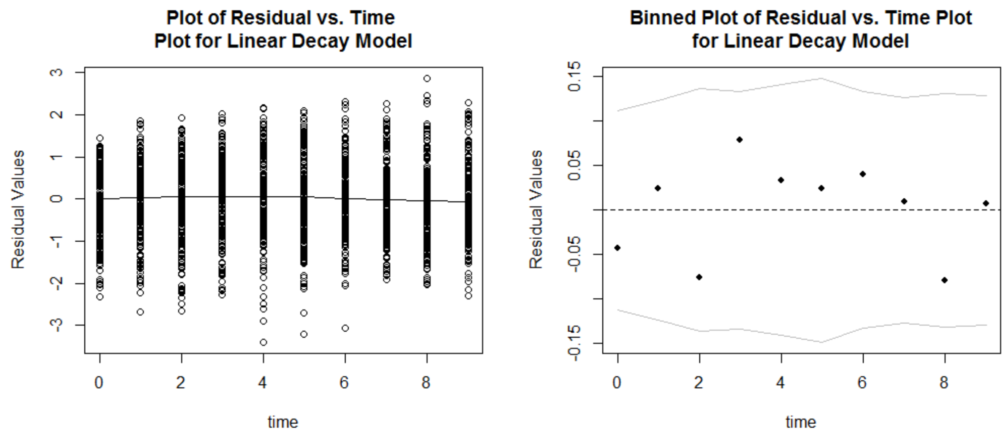


Figure 12: *Residuals versus time and binned residuals vs. time*

4.3 Model Comparison

To choose the best way to model the log odds decay, we study the linear decay as seen in the first example, the exponential decay, and the quadratic decay. We will study their results to understand how their decay looks, which model fits the data best, and whether there are any other model-checking problems that arise. Before moving to model checking, we note that the Bayesian diagnostics showed strong results with convergence and no autocorrelation.

The first way we will check the models is to look at the decay that each three log odd decay models create. We do this by simulating random values of the covariates and calculating log odds based on the posterior means of our parameters and the set simulated coefficients (used to simulate the data sets). We then convert the log odds into probabilities, which allows us to see the sort of decay that each of the three models calculates and how well the posterior distribution matches its simulated curve. Figure 13 displays the results of the calculation using data values of an 8-year old, male student from high.economic category family. We see in the plot that the linear decay on the log odds scale matches its simulated values the best of the three different decay curves. Overall, the linear decay on the log odds scale shows a decline over time, but the probability that a student is not on task never reaches zero, and even at the end of the time period about 50% of students are on task. Whereas the quadratic decay on the log odds scale shows a steeper decline, pronouncing the time effect on a student' attention allocation. Additionally, we see that our posterior distribution creates a shallower curve, declining less drastically than the simulated values. Lastly, the exponential curves seem to show the worst matching between the simulated values and the fitted posterior distribution. We see that there is a much shallower decline over time, with predicted on task behavior staying under 50% for the majority of the curve. This displays provides a better understanding of whether our models are working and what effect each type of log odd decay has on the decline of attention.

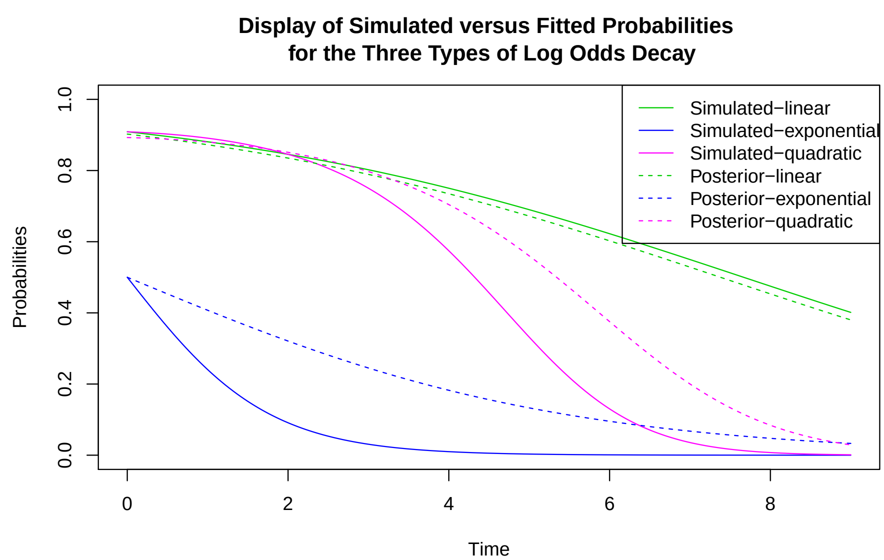


Figure 13: *Comparison plot of three models with simulated values and fitted values*

Next, we will look to model fitting criterions to better understand how well the

Table 2: DIC Criterion

Linear	Exponential	Quadratic
5479.0	6139.8	5074.0

models are fitting the data. In Table 2, we see the three values for the DIC criterion for each of the declines on the log odds scale. We look to the lowest DIC value, which is the Quadratic model by a large difference. Before making conclusions that the quadratic model is the best model to move forward with, we should also check the quadratic model's residuals.

From Figure 14, we can see that the residuals are not randomly scattered around zero, showing evidence of assumption violations of the errors. We see more closely in the binned residual plot that as time increases, the errors are more positive meaning that the model under fits the data during those times (specifically between 4 and 8 on the time scale). This shows us evidence that the model does not do a good job at fitting the data. Despite the lower criterion value, we do not continue to use the quadratic model but instead move forward using the linear log odds decay model.

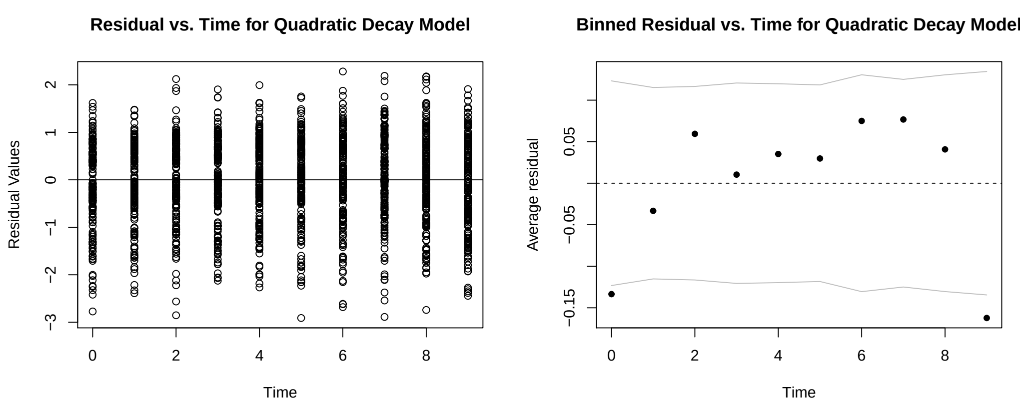


Figure 14: *Plots of Residuals versus Time for Quadratic Log Odds Decay Model*

5 Classroom Study Data

5.1 Model Building with Data

To begin model building with the classroom data, we first start with looking at the results as a binomial outcome, on task or off task. We analyze the structure of the data and the variables provided that would help predict whether or not the student is on task. Due to the hierarchical structure of the data, we decided to begin by making a model to reflect the nested relationships of classrooms and students. We create random intercepts for both the 745 students and the 22 classrooms. Before beginning the Bayesian analysis, we use a frequentist approach to the hierarchical model using the R package, lme4. Within the model we model random intercepts for students and classrooms and include the different activities, grades, and time as fixed effects. We see the following results shown in Table 3.

Table 3: LMER output for mixed effects model

	Estimate	SD	P-value
β_0	-0.78	0.06	$< 2e^{-16}$
β_{fem}	-0.23	0.06	0.00
β_{other}	-0.46	0.28	0.11
$\beta_{sgindiv}$	0.08	0.08	0.36
β_{indiv}	0.10	0.07	0.13
$\beta_{sgteach}$	-0.43	0.13	0.00
$\beta_{wholecarpet}$	0.28	0.07	0.05
$\beta_{five.min}$	0.02	0.01	0.03

Overall the model shows that female, sgteach, wholecarpet, and five.min are the most significant predictors of on task behavior. Also, on-task behavior for the other and small group teach is significantly less likely than for the whole desk baseline category. This gives us a strong starting point to check our results when performing the Bayesian analysis, shown in Equation 3.

$$
\begin{aligned}
y_i \sim\ & \text{Binomial}(\text{inverse.logit}(\alpha_{0ji} + \alpha_{0ki} + \beta_{femalei}\text{female}_i + \beta_{indivi}\text{indiv}_i \\
& +\beta_{sgindivi}\text{sgindiv}_i + \beta_{sgteachi}\text{sgteach}_i + \beta_{whocarpetsi}\text{wholecarpets}_i \\
& +\beta_{five.mini}\text{five.min}_i \qquad (3)
\end{aligned}
$$

$$
\begin{aligned}
\alpha_{0j[i]} &\sim \text{Normal}(\beta_0, \tau^2) \\
\alpha_{0k[i]} &\sim \text{Normal}(\beta_0, \tau_2^2) \\
\beta_{femalei} &\sim \text{Normal}(0, 1000) \\
\beta_{indivi} &\sim \text{Normal}(0, 1000) \\
\beta_{otheri} &\sim \text{Normal}(0, 1000) \\
\beta_{sgteachi} &\sim \text{Normal}(0, 1000) \\
\beta_{sgindivi} &\sim \text{Normal}(0, 1000) \\
\beta_{wholecarpetsi} &\sim \text{Normal}(0, 1000) \\
\beta_{five.mini} &\sim \text{Normal}(0, 1000) \\
\\
\tau &\sim \text{Uniform}(0, 10) \\
\tau_2 &\sim \text{Uniform}(0, 10)
\end{aligned}
$$

Within the model we use the variables, gender and activity. Activity is coded using dummy variables, with wholedesk as the baseline because it had the largest sample size compared to the other activities, allowing the model to be more stable in the calculations. Additionally in the model are the variable five.min, derived from the data sets time variable which provides the hour, minute, and second of the time the observation was taken. We created the five.min variable to represent the time in five minute increments since the start of a session and an activity. Therefore, if during a session the teacher changes activity, we restart the time during the start of that activity. One of our goals is to answer how well students stay on task when proposed with new activities, and since we believe that students will regain attention at the beginning of each new activity we adjust the time variable to measure the time since the beginning of a new activity and session.

We decide to use the priors we used in the simulated models, with a distribution of normal centered at 0 with precision 0.0001. Using research by Andrew Gelman, we decide to use a uniform between 0 and 10 for the variance parameter of the random intercepts, being that 10 is a relatively unreachable upper bound in our logistic model [1]. This sets for a weakly informative variance parameter, which should allow for the likelihood to match frequentist methods in the posterior distribution.

To check this, we compare our Bayesian fits with frequentist mixed effects models using the R package, lme4. In Table 4, we see that the coefficient estimates from the posterior mean are very similar to those seen in the lmer output from Table 3. This reassures that our model is effective in the binomial case. However, the data provide a variable that explains the students behavior when off-task. Next, our model moves towards the multinomial case where we study the effect of time, activity, and gender on the students behavior when off task.

Table 4: Posterior output for Bayesian mixed effects model

	Estimate	SD	MCMC Error	Credible Interval
β_0	-0.78	0.03	0.003	$(-0.34, -0.55)$
β_{fem}	-0.23	0.06	0.004	$(-0.35, -0.13)$
β_{other}	-0.48	0.30	0.005	$(-1.12, 0.08)$
$\beta_{sgindiv}$	0.08	0.09	0.004	$(-0.09, 0.23)$
β_{indiv}	0.10	0.07	0.004	$(-0.02, 0.25)$
$\beta_{sgteach}$	-0.43	0.13	1.57	$(-0.68, -0.17)$
$\beta_{wholecarpet}$	0.28	0.08	0.02	$(0.14, 0.44)$
$\beta_{five.min}$	-0.003	0.01	0.003	$(-0.12, 0.01)$

5.2 Multinomial Model with Simulated Response Data

Our overall goal for the project is to better understand students attention allocation in the classroom and what distracts them during different lessons. Up until now, we looked at the data in a binomial viewpoint, understanding simply on or off task. However, we now want to understand better why they are off task. This is done by using the behavior variable, which labels the student off task with different activities. Based on the biplot shown in Figure 15, we see that there are three choices of behavior that maximizes the variance of the data: on task, peer, and a category that encompasses the other categories, named in our case, other. We will fit a multinomial model to predict whether the students are on task, peer, or other predicted with the covariates female, activity, and time, and random intercepts for student.

Before using the full dataset, we simulate a response variable based on the data's covariates. The response, y, is simulated as a multinomial variable with three options, 1=onTask, 2=Peer, 3=Other, controlled with the covariates: wholedesks, female, five.min, and random intercepts per student displayed in Equation (4).

$$\begin{aligned} \log\left(\frac{p_j}{1-p_{OT}}\right) &= \alpha_{0ji} + \beta_{femalej}\text{fem}_i + \beta_{wholedesksj}\text{wholedesks}_i \\ &\quad + \beta_{five.minj}\text{five.min}_i + \text{true.RI}_j \sim \text{Normal}(0,\tau^2) \end{aligned} \tag{4}$$

Where j=o for other, and p for peer, and i=1...N, where N is number of subjects.

We chose the values of the coefficients during the simulation to be $\beta_{0.p} = 0.4$, $\beta_{0.o} = 0.6$, $\beta_{female.p} = 0.5$, $\beta_{female.o} = -0.5$, $\beta_{wholedesks.p} = -0.2$, $\beta_{wholedesks.o} = -0.2$, $\beta_{five.min.p} = -0.5$, $\beta_{five.min.o} = -0.5$ (the letters p and o standing for the peer choice and other choice for the response).

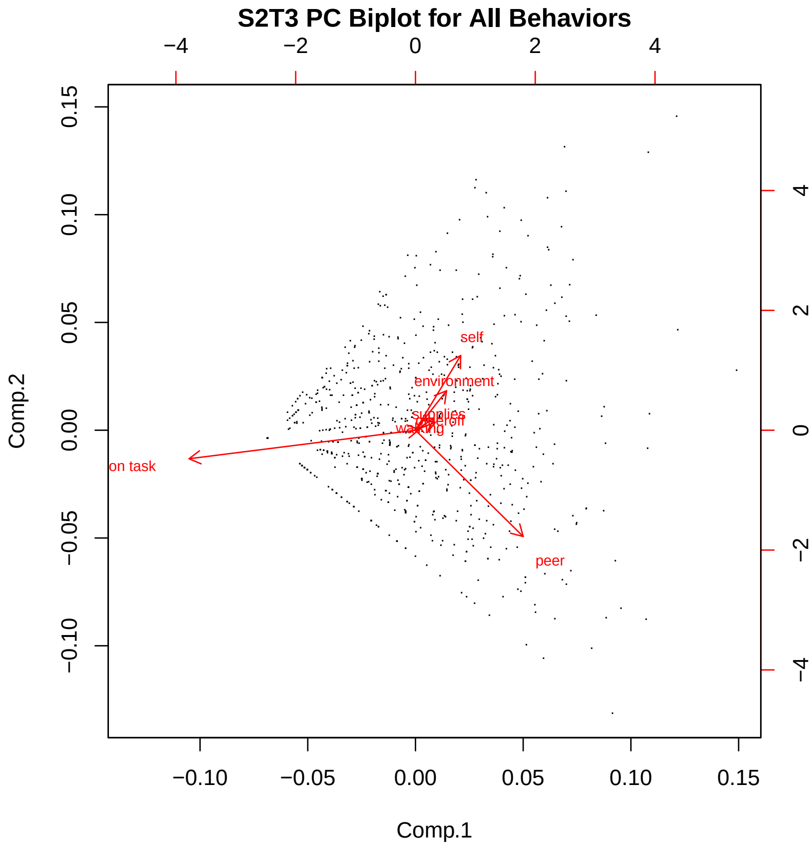


Figure 15: *Biplot of principle component analysis on behavior*

The next step is to model this data using a Bayesian multinomial model. There is no option for checking this in a Frequentist method using lme4 or R packages, mlogit or nnet. Additionally, because the random intercepts per students have different sample sizes some of which being quite small, a Bayesian method will be unaffected by the small sample sizes and will predict better than any Frequentist approach [6]. The multinomial equation in hierarchical Bayes form and WinBugs code form is shown in Equation (5).

$$
\begin{aligned}
y_i \sim\ & \text{Multinomial}(\text{inverse.logit}(\alpha_{0ji} + \alpha_{0ki} + \beta_{femalei}\text{female}_i \beta_{whocarpetsi}\text{wholecarpets}_i \\
& + \beta_{five.mini}\text{five.min}_i
\end{aligned} \tag{5}
$$

$$
\begin{aligned}
\alpha_{0j[i]} &\sim \text{Normal}(\beta_0, \tau^2) \\
\alpha_{0k[i]} &\sim \text{Normal}(\beta_0, \tau_2^2) \\
\beta_{femalei} &\sim \text{Normal}(0, 100) \\
\beta_{wholecarpetsi} &\sim \text{Normal}(0, 100) \\
\beta_{five.mini} &\sim \text{Normal}(0, 100) \\
\\
\tau &\sim \text{Uniform}(0, 4) \\
\tau_2 &\sim \text{Uniform}(0, 4)
\end{aligned}
$$

Rube Code:

```
multinom.model<- "model{
        for(i in 1:NOBS){
                mu[i, 1] <- bp0+ LC(bp, COV, , i)
                        + bfive.minp*five.min[i] + rip[index[i]]
                mu[i, 2] <- bo0+ LC(bo, COV, , i)
                        + bfive.mino*five.min[i] + rio[index[i]]
                emu[i,1] <- 1
                emu[i,2] <- exp(mu[i,1])
                emu[i,3] <- exp(mu[i,2])
                for(j in 1:3){
                        p[i,j] <- emu[i,j]/sum(emu[i,1:3])
                }
        y[i]~ dcat(p[i,1:3])
        }


###Priors###
        bfive.minp~dnorm(0, BT.PREC=1e-4)
        bfive.mino~dnorm(0, BT.PREC=1e-4)
        FOR(bp, COV, , ?~dnorm(0, COV.PREC=1e-4))
        FOR(bo, COV, , ?~dnorm(0, COV.PREC=1e-4))
        bp0~dnorm(0, 1e-4)
        bo0~dnorm(0, 1e-4)
        ri.precp <- pow(ri.sdp, -2)
        ri.sdp ~ dunif(0, RI.SD.MAX=5)
```

```
        ri.preco <- pow(ri.sdo, -2)
        ri.sdo ~ dunif(0, RI.SD.MAX=5)
        for(i in 1:NS){
                rip[i]~ dnorm(0, ri.precp)
                rio[i]~ dnorm(0, ri.preco)
        }
}"
```

We kept the priors consistent from the binomial analysis, to allow the likelihood to have the larger affect on the posterior than the prior distributions. You will notice in the Rube Code that the y is distributed from the "dcat" command which chooses a category of y based on the probability vectors calculated using the log odds of peer, other, and the probability of on task which is just 1 minus the other two probability vectors. The results are shown in Table 5. We can check that the model code worked well and accurately fitted the model by checking the simulated coefficient values to the posterior means.

Table 5: Posterior distribution of multinomial model using simulated data

	Mean	SD	MCMC error	Credible Interval	Simulated Value
$\beta_0 p$	0.45	0.05	0.002	$(0.36, 0.55)$	0.40
$\beta_0 o$	0.72	0.05	0.002	$(0.63, 0.82)$	0.60
$\beta_{pfemale}$	0.51	0.05	0.001	$(0.41, 0.61)$	0.50
$\beta_{ofemale}$	-0.52	0.06	0.002	$(-0.63, -0.41)$	-0.50
$\beta_{pwholedesks}$	-0.20	0.05	0.001	$(-0.31, -0.09)$	-0.20
$\beta_{owholedesks}$	-0.24	0.06	0.002	$(-0.36, -0, 13)$	-0.20
$\beta_{five.minutep}$	-0.52	0.05	0.002	$(-0.55, -0.49)$	-0.50
$\beta_{five.minuteo}$	-0.53	0.05	0.002	$(-0.56, -0.60)$	-0.50
$\beta_{ri.sdp}$	0.06	0.05	2.35	$(0.003, 0.19)$	0.10
$\beta_{ri.sdo}$	0.72	0.05	0.007	$(0.006, 0.18)$	0.10
$\beta_{rip[1]}$	0.006	0.07	0.007	$(-0.144, 0.16)$	
$\beta_{rio[1]}$	-0.001	0.09	0.003	$(-0.21, 0.19)$	

We see from Table 5, that the model did a very good job at fitting the data, marked by matching posterior means to the simulated coefficients, with all of the true values inside the credible intervals except for $\beta_0 o$ being slightly outside the 95%

interval. The only concern is the large MCMC error in the ri.sdp, which could be a result of not enough iterations. Overall the diagnostics for the posterior distribution showed no autocorrelation amongst the iterations and strong convergence in every variable (with the exception of ri.sdp). These results prove that the model fits multinomial, nested data well and will be appropriate for use on the true study data.

5.3 Classroom Study Results- Multinomial Logistic Model with Random Intercepts per Student

Finally, we begin to study the data from the classroom study to understand what causes students to be off task in the classroom. We use the multinomial mixed effects model described in Section 5.2 to analyze the study data. We continue to model random intercepts per student, but will also explore random intercepts per classroom. The majority of this paper analyzed the methodology of the modeling and described the Bayesian principles to support the model we use with the real data. Now we will shift gears and understand the results and what this means to Dr. Anna Fisher's study on student attention allocation in the classroom.

First, we will begin by checking the diagnostics of the posterior distribution. summarizing the results in Table 6, showing the posterior mean, standard deviation, MCMC error and credible intervals. The covariates, whole carpet and individual, are added into the model describing what activity the students are performing while being observed. These are included as dummy variables. We choose whole desks, whole carpets, and individual as the three activities in the model because they contain the most observations, leaving the baseline a combination of the remaining activities. In Figures 16-21, we show that all parameters converged with Rhat values around 1 and very little autocorrelation amongst iterations. Additionally the plots show us the credible interval plots to get an understanding which coefficient means are significant versus those that are not. We check this by looking for 0 in the interval shown in the density plot for each figure. The diagnostics that these plots show are critical before moving forward and interpreting the results.

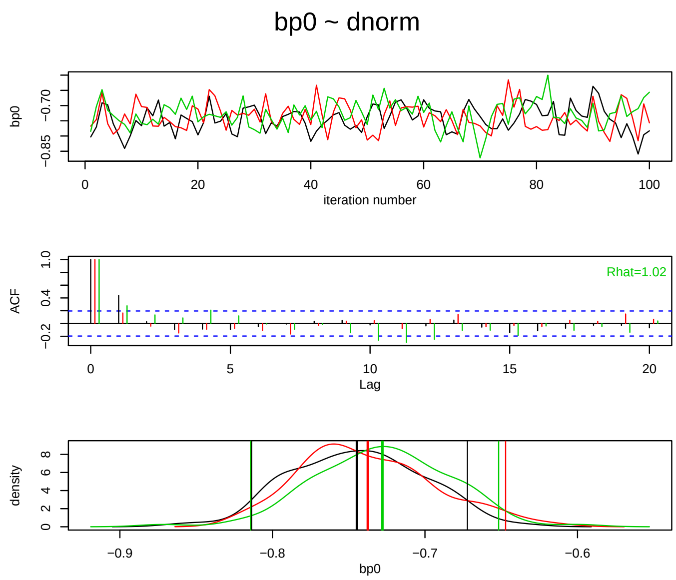


Figure 16: *Posterior mean of peer intercept*

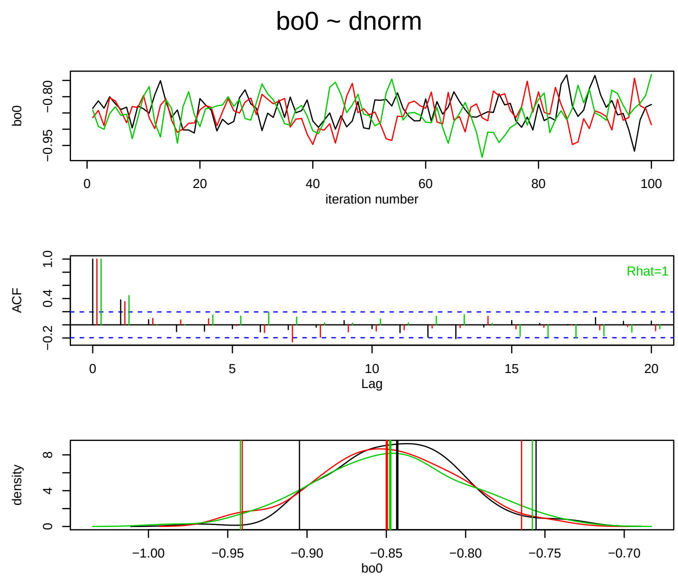


Figure 17: *Posterior mean of other intercept*

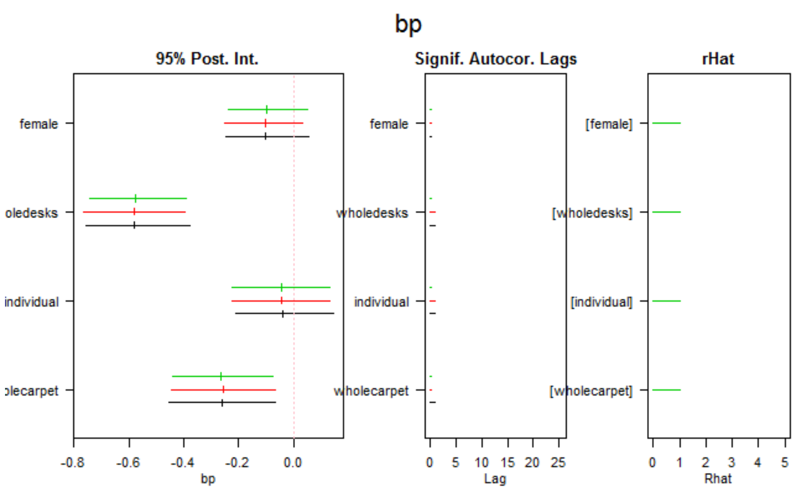


Figure 18: *Posterior mean of peer covariates*

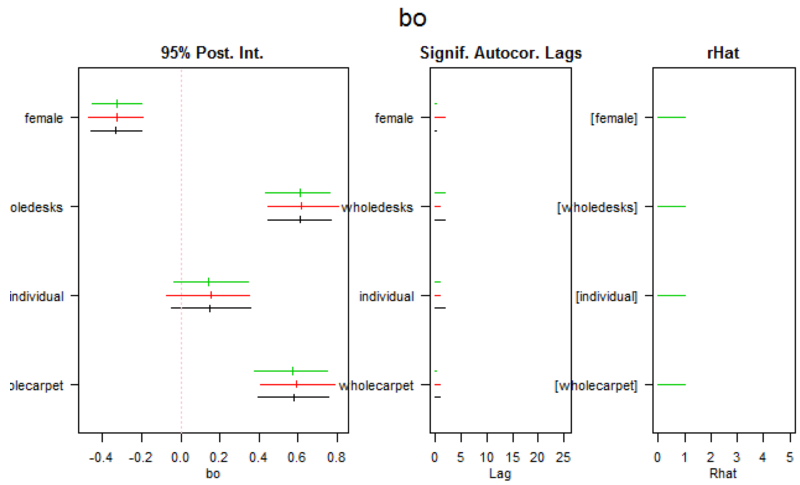


Figure 19: *Posterior mean of other covariates*

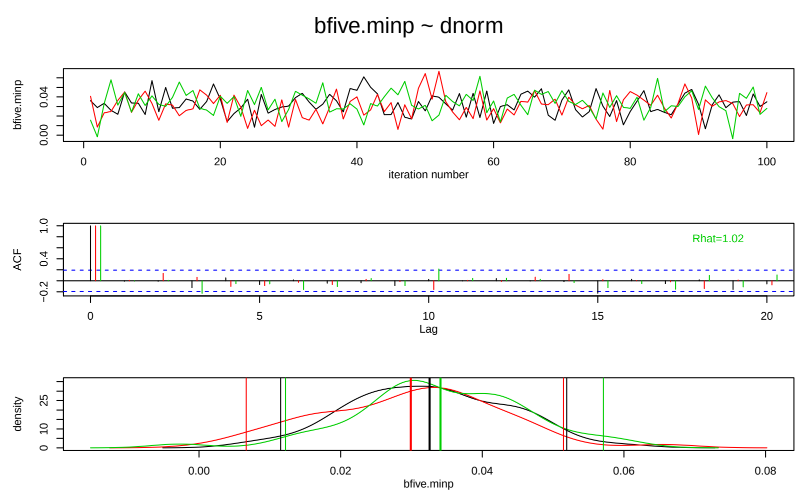


Figure 20: *Posterior mean of peer five minute variable*

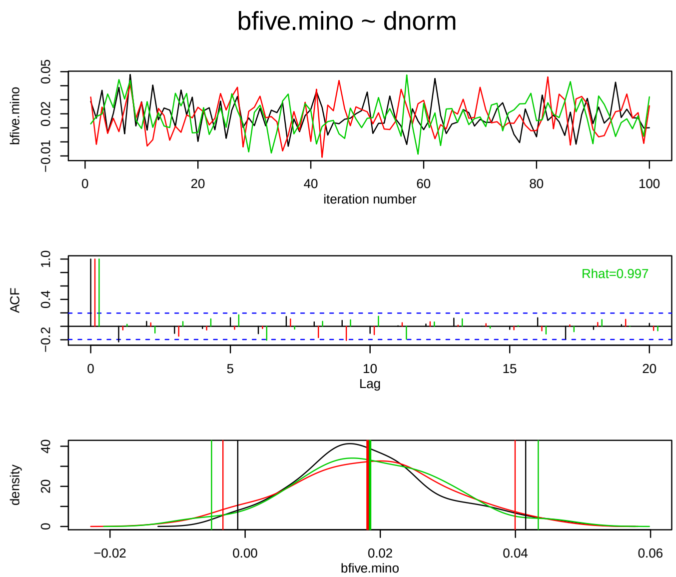


Figure 21: *Posterior mean of other five.minute variable*

Table 6: Posterior distribution of multinomial model using classroom study data

	Mean	SD	MCMC error	Credible Interval
β_{0p}	-1.48	0.09	0.003	$(-1.66, -1.31)$
β_{0o}	-1.69	0.09	0.002	$(-1.86, -1.52)$
$\beta_{pfemale}$	-0.10	0.07	0.005	$(-0.25, 0.05)$
$\beta_{ofemale}$	-0.33	0.07	0.005	$(-0.46, -0.20)$
$\beta_{pwholedesks}$	-0.58	0.09	0.007	$(-0.76, -0.39)$
$\beta_{owholedesks}$	0.61	0.09	0.007	$(0.44, 0.78)$
$\beta_{pwholecarpet}$	-0.26	0.10	0.007	$(-0.45, -0.07)$
$\beta_{owholecarpet}$	0.58	0.10	0.007	$(-0.40, 0.77)$
$\beta_{pindividual}$	-0.04	0.09	0.007	$(-0.22, 0.13)$
$\beta_{oindividual}$	0.15	0.10	0.007	$(-0.05, 0.35)$
$\beta_{five.minutep}$	0.03	0.01	0.001	$(0.009, 0.06)$
$\beta_{five.minuteo}$	0.02	0.01	0.001	$(-0.005, 0.04)$
$\beta_{ri.sdp}$	0.64	0.04	3.93	$(0.56, 0.73)$
$\beta_{ri.sdo}$	0.64	0.04	0.003	$(0.57, 0.71)$
$\beta_{rip[141]}$	1.61	0.43	0.004	$(0.36, 2.01)$
$\beta_{rio[1]}$	1.02	0.40	0.03	$(0.26, 1.78)$

Next, we look at the results summarized in Table 6. The coefficients are labeled with either p or o to mark whether they describe off task with peer or off task with other for the response variable. In multinomial logistic regression, we calculate two sets of coefficients to represent the relationship between peer and on task and other and on task, similar to calculating two separate regressions [3]. The intercept for the peer response shows us that when the student is a male and doing an activity that is not whole desks, whole carpet, or individual, and the time of the observation is 0, the odds for this baseline subject of peer off task behavior versus on-task is ($e^{-1.48}$) = 0.23, indicating that on task behavior is 1/0.23 = 4.3 times as likely as off-task behavior. The odds of other off task behavior versus on task behavior is $e^{-1.69} = 0.18$ or the inverse of that shows on task behavior to be 5.4 times as likely as other off-task behavior. Together we have a ratio of on task to peer to other of $1 : 0.23 : 0.18$ which corresponds to $71\% : 16\% : 13\%$ for the three behaviors when the other covariates are held constant at the baseline.

Because both the five.min posterior means are positive, we know that with time there is a stronger chance of being off task with peer or off task with other compared to being on task. The strong negative intercepts show us that at time 0, or the beginning of the session or activity, a student has greater odds of being on task than off task. For an additional five minutes of time, the odds of being off task behavior with peer versus on task is ($e^{0.03}$)=1.03 times great than the ratio of the peer versus on task at baseline 0. Similarly for off task with other, the odds of a baseline student observed five minutes into the activity (or session) being off task with other versus on task is 1.02, describing that five minutes into the study, the odds ratio between other off task to on task is 1.02 times the odds of being off task versus on task at the beginning of the activity or session.

The different activities show to be interesting factors in a students' behavior in the classroom. The odds of being off task with peers vs. on task when the activity is "whole desks" is 0.56 times the odds of being off task with peers vs. on task when the activity is other (not whole desk or whole carpet or individual). The odds of a student being off task with other versus on task during the activity whole desks are 1.84 times greater than the odds of being off task with other versus on task when the activity is other. This leads us to understand that during the activity whole desks a student is more likely to be off task with others than on task or off task with a peer. This pattern follows for whole carpet and individual, showing that students tend to be off task with either self-distraction, environmental distraction, sleeping, moving around the classroom, or other than being off task with a peer or being on task. Additionally, a student that is female has lower odds than a male of being off task with either peer or other.

We notice in Figure 22 and Figure 23 that most of the student intercepts' credible intervals cross 0, meaning they do not have a strong impact per student at explaining what behavior they are most likely in the classroom. This leads us to believe that modeling a random effect per student may not be the best option, and we continue by looking at a random intercept per classroom. Before we look to a different model, we check the prior posterior plots to assure that the priors chosen are appropriately diffuse. In Figure 24, we see that the posterior distributions on the random intercepts show a much higher density compared to the prior signifying that the prior is not influencing the posteriors greatly, but instead is serving as a non-informative prior.

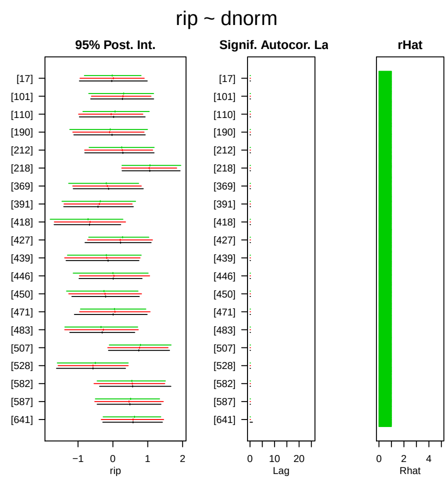


Figure 22: *Posterior mean of all random intercepts for peer*

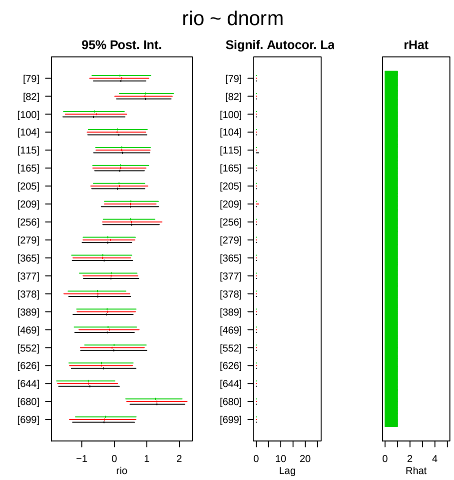


Figure 23: *Posterior mean of all random intercepts for other*

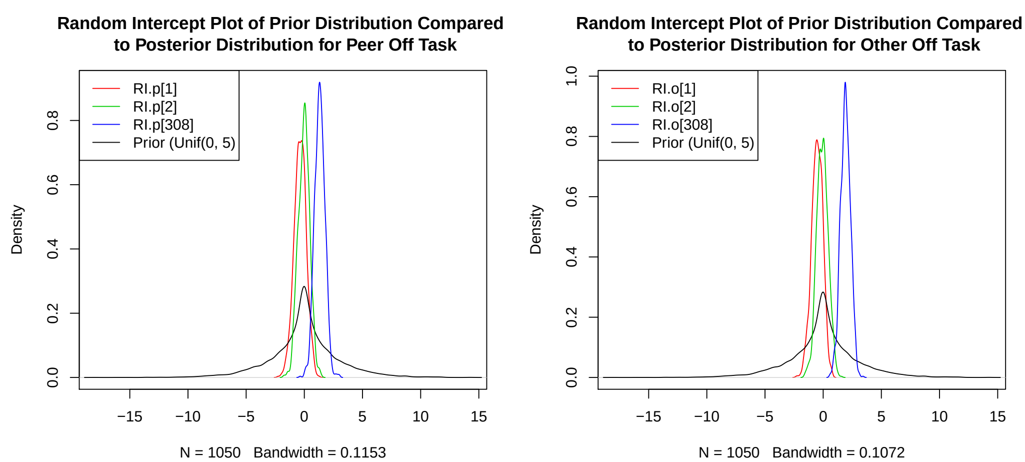


Figure 24: *Prior/posterior plots of student random intercepts*

5.4 Results: Classroom Random Intercept Model

We move forward by exploring a model with random intercept per classroom instead of student. The DIC of the student random intercept model was 18796.2 compared to the classroom random intercept model's DIC value of 18692.1. The difference is quite significant and leads us to continue further analysis using the classroom RI model. In Table 7, we see that the values of the coefficients do not change much. The same coefficients show to be significant, β_{0p}, β_{0o}, $\beta_{femaleo}$, $\beta_{wholedesksp}$, $\beta_{wholedeskso}$, $\beta_{five.minp}$, with the exception of $\beta_{wholecarpetp}$ no longer significant. We also see a change in sign for some of the activities' posterior means for the peer off task response. Both whole carpet and individual have positive off task with peer posterior means, different from their negative coefficients in the student random intercept model. This shows that when creating random intercepts for classroom instead of students, there is not an opposite effect for the activities, whole carpet and individual, for off task with peer and off task with other. Additionally, we now see a stronger affect of the random intercept per classrooms. This could be due to the fact that each group consists of more data and therefore provides more information than just a single students observations. In Figure 25-26, we see the more pronounced effects of the classroom random intercepts.

Table 7: Posterior distribution of multinomial model using classroom study data

	Mean	SD	MCMC error	Credible Interval
β_{0p}	-1.53	0.13	0.006	$(-1.80, -1.28)$
β_{0o}	-1.59	0.12	0.008	$(-1.84, -1.38)$
$\beta_{pfemale}$	-0.10	0.05	0.002	$(-0.21, 0.002)$
$\beta_{ofemale}$	-0.33	0.05	0.002	$(-0.42, -0.23)$
$\beta_{pwholedesks}$	-0.52	0.10	0.004	$(-0.71, -0.32)$
$\beta_{owholedesks}$	0.54	0.10	0.004	$(0.34, 0.74)$
$\beta_{pwholecarpet}$	0.005	0.12	0.004	$(-0.23, 0.23)$
$\beta_{owholecarpet}$	0.60	0.11	0.004	$(-0.40, 0.80)$
$\beta_{pindividual}$	0.10	0.10	0.004	$(-0.12, 0.32)$
$\beta_{oindividual}$	0.18	0.11	0.005	$(-0.02, 0.40)$
$\beta_{five.minutep}$	0.03	0.01	0.001	$(0.009, 0.06)$
$\beta_{five.minuteo}$	0.01	0.01	0.0004	$(-0.01, 0.04)$
$\beta_{ri.sdp}$	0.47	0.09	0.47	$(0.32, 0.67)$
$\beta_{ri.sdo}$	0.38	0.07	0.003	$(0.27, 0.54)$
$\beta_{rip[1]}$	0.14	0.15	0.003	$(-0.20, 0.43)$
$\beta_{rio[1]}$	-0.22	0.15	0.008	$(-0.52, 0.05)$

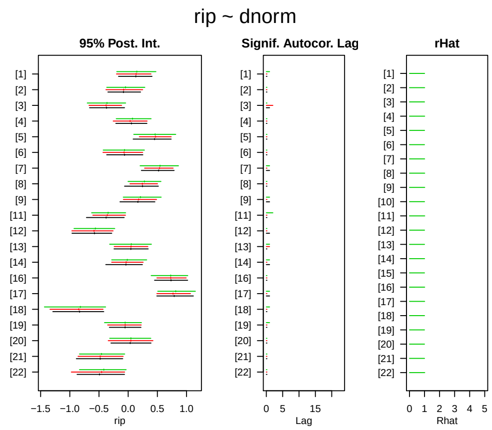


Figure 25: *Posterior mean of all classroom random intercepts for peer*

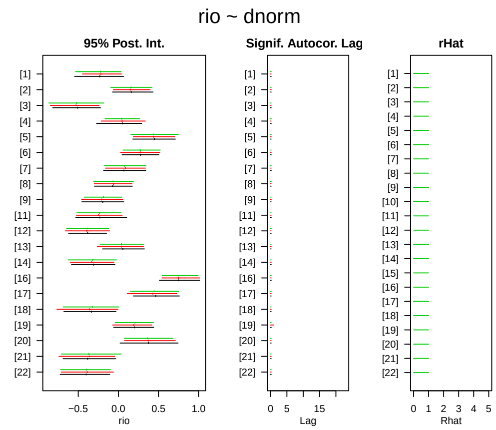


Figure 26: *Posterior mean of all classroom random intercepts for other*

Lastly, we look to residuals to assure the model fit the data well. In Figures 27 and 28, we see two plots of residuals versus time with a lowess smoother line overlaying the points for both off task with peer and off task with other responses. The residuals are calculated for peer off task by subtracting the probability of being off task with peer from a 1 or 0 corresponding with whether the student is off task with peer or not. This creates the separation in the residuals that we see in the plot. We see however that the lowess smoother appropriately models the residuals, showing a flat line depicting the fact that there is no further trend over time that the model did not recognize. We see a very similar result for other off task behavior.

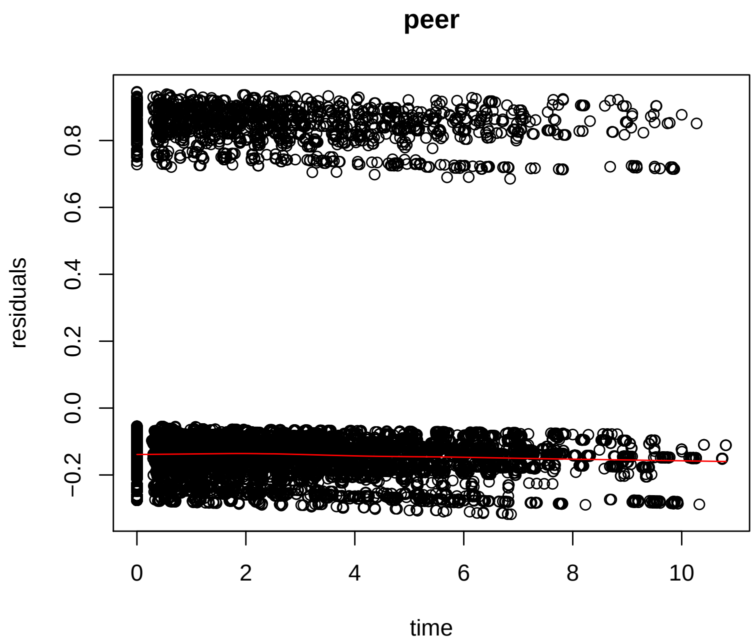


Figure 27: *Residuals versus Time for Peer Off Task Response*

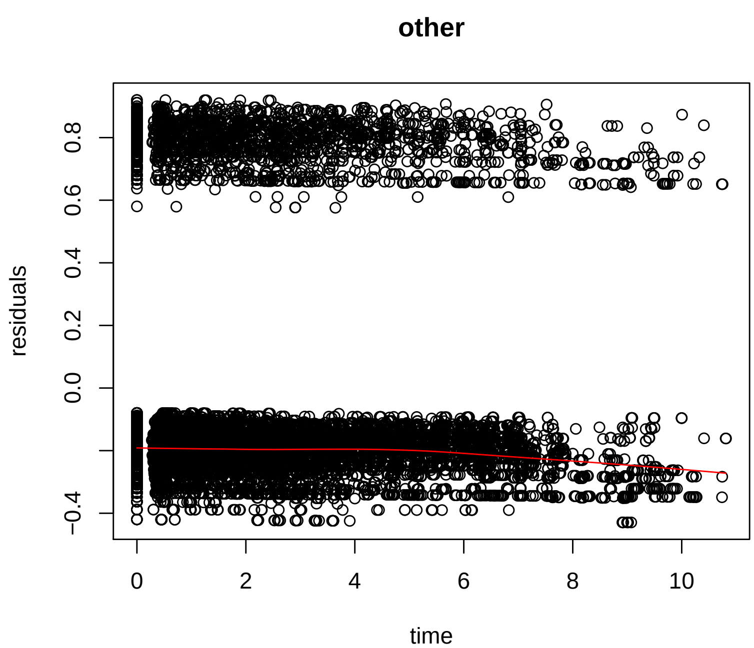


Figure 28: *Residuals versus Time for Other Off Task Response*

Other ways to visualize and understand our results are to compare the probabilities of being on task, off task with a peer, or off task with other using our posterior means and different values of the covariates. By plotting these together on a ternary plot, we are able to better understand how a different activity or gender affects their probability to be on task, off task with peer, or off task with other. In Figure 29, we see the ternary plot of the probabilities of the different combinations of activities at time 0. The plot illustrates that when all other covariates are held constant, activity affects a students attention allocation between on task, peer, and other. We see that the results of the plot are similar to what we see in the posterior means. Wholedesks is the red point that is furthest towards the other/on task axis, meaning it is least likely to be off task with peer. We see that other and individual are very close together, depicting that they have similar effects on a students' attention. Wholecarpet is the point furthest from on task corner, meaning that it has the lowest probability of being on task compared to the other three activities. Additionally, we are able to see whether time has an effect on a students attention allocation. Figure 31provides evidence that over time students are more likely to be off task, we see this by the short trail of red points on the ternary plot. The plot is representing a student during the activity whole desks and shows that over 25 minutes (five.min=1:5) the student moves away from the on task corner. The trail is quite short, showing that there is not a large time effect, which is evident in the small posterior mean value.

These results explain important aspects of the data, helping us to understand the different ways students' attention changes due to activities and time.

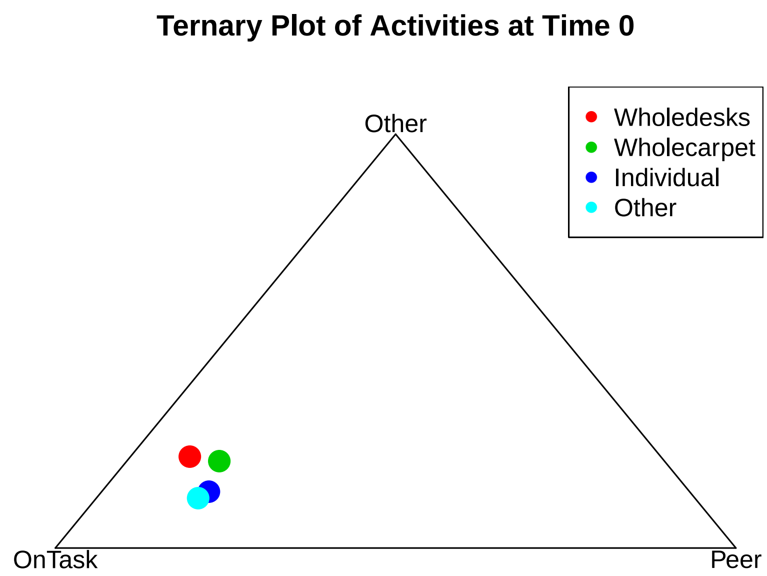


Figure 29: *Ternary plot of probabilities of response given different activity and gender effects*

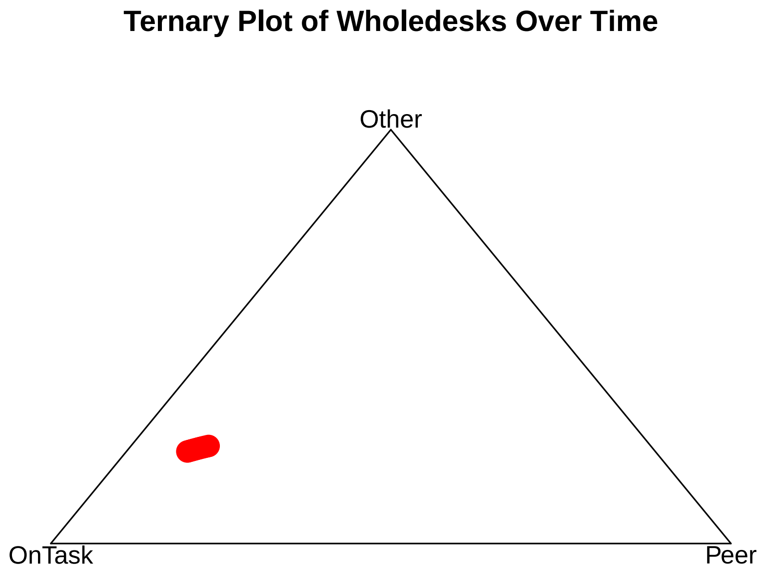


Figure 30: *Ternary plot of probabilities of response given different activity and gender effects*

6 Conclusion

As we conclude, it is important to note the effectiveness and efficiency of WinBUGS and the R package, rube, to run advanced models, not available in standard software. We were able to test our scientific hypothesis of whether time was a decaying factor to describe a student's behavior, and what activities contributed to a student's attention allocation. Because running multinomial, mixed effects logistic regressions is not easily available in most software packages, we were able to perform the analysis using weak priors, and the quick performance of WinBUGS. The analysis shows the interchangeability between Bayesian analysis and frequentist approaches when informative priors are not available and non-informative priors are used.

We began our analysis by first using completely simulated data with fictional covariates to understand the use of WinBUGS, rube, and the Bayesian output. Bayesian model fitting differs quite greatly from Frequentists' approach. In Bayesian analysis, we must check for the accuracy of the MCMC chains and the effectiveness of our priors. Through the simulations, we were able to test different hypothesis for attention decay through binomial models and choose a best form for attention decays and initial priors. When moving to the classroom study data, we continued with the diffuse priors and predicting a binomial response. By predicting binomial responses, we were able to compare the results of an lmer output to our Bayesian analysis to check for accuracy in the model. With confidence in the binomial model using the real classroom data, we moved forward to our final models. Because little background information was provided from experts, we did not model our data using informative priors and therefore did not want to see a posterior distribution greatly influenced by the prior distributions. This created results very similar to the results from Frequentists approaches (when possible to check with R packages), and allowed us the flexibility to model further when software was not provided for more complex models.

Our analysis of the student attention allocation classroom study produces interesting results of how the classroom environment affects students. The analysis shows that students are more likely to be on task at the beginning of a session or activity and over time slowly move to being off task. Additionally, we see the gender effect come into play for elementary school students, with more females staying on task than males. Lastly, activities in the classroom tend to affect students' attention allocation. The activities, whole carpet, individual, and other, sway students' attention towards off task with other and off task with peer rather than on task. However, whole desks shows that a student' attention moves to being off task with other more so than on task or off task with peer. These results are a good starting point for the study, and with further data collected and more information provided, further conclusions can be made. Future information that would be interesting to note is the environment surrounding the students, and whether pictures on the walls or brightly colored posters become distractions for students. The study plans to collect data on the environmental factors of the actual classroom, which could become a very interesting predictor of the student's attention behavior.

In future work, I would like to move forward with this analysis with a further investigation of the underlying assumptions of our model and expanding the model for deeper understanding. For instance, our model assumes that a student's attention is more focused at the start of an activity and session. We implemented this by restarting the five.min time variable as zero at the beginning of each activity and session. Additionally, we modeled an overall decay for all activities and sessions. A next step to further investigate these assumptions could be to learn more about the interaction between activity and time and see whether each activity should have its own time decay on the log odds scale. Lastly, an interesting feature in the model could be time of day, and whether morning versus afternoon sessions affect students. These aspects of the model would be important to accomplish moving forward in future work. With further information on student's classroom behavior from an expert, we would be able to calculate posterior distributions that contain more information than just from the data, but also scientifically proven facts implemented using prior distributions. Additionally, further steps could be taken in the modeling, looking at more random effects, interactions, and perhaps random slopes based on time. Because this was a process of learning Bayesian analysis as well as experimenting with models in WinBUGS with little documentation for reference, we did not exhaust all model checking possibilities, and further work could be addressing more details in the model and model checking using more advanced techniques.

7 References

1. Gelman, Andrew. "Prior Distributions for Variance parameters in Hierarchical Models (Comment on Article by Browne and Draper)." Bayesian Analysis 1.3 (2006): 515-534.

2. Gelman, Andrew, and Jennifer Hill. Data analysis using regression and multilevel/hierarchical models. Cambridge: Cambridge University Press, 2007.

3. Hedeker, Donald . "A Mixed-Effects Multinomial Logistic Regression Model." statistics in medicine 22 (2003): 1433-1446.

4. "Institute for Digital Research and Education." Annotated Stata Output: Multinomial Logistic Regression. UCLA: Statistical Consulting Group, n.d. Web. 15 Apr. 2013. $< http : //www.ats.ucla.edu/stat/stata/output/stata_m logit_o utput.htm >$.

5. John W. Seaman III, John W. Seaman Jr. & James D. Stamey (2012): Hidden Dangers of Specifying Noninformative Priors, The American Statistician, 66:2, 77-84.

6. Starkweather, Jon, and Amanda Kay Moske. "Multinomial Logistic Regression." (2011).