Comparing models for predicting high school dropout for rising ninth graders in rural North Carolina

Emily Hadley

April 22, 2015

Abstract

This study compares the accuracy and strength of logistic regression, multilevel logistic modelling, principle component analysis (PCA), and Bayesian hierarchical modeling to predict high school dropout in rural North Carolina. The most significant predictor of high school dropout in the logistic regression model and the mulilevel logistic regression model is achievement in math, but both models struggled with poor model fit. PCA was the most accurate predictor of high school dropout, but does not allow for interpretation of coefficients. The Bayesian modelling shows promise, but the model did not converge due to computational limitations. Policymakers looking to identify at-risk rural students should consider using more complete multilevel logistic regressions, only focusing on prediction with PCA, or solving computational issues from Bayes.

Keywords: education, rural, dropout, Bayes, PCA, logistic regression

1 Introduction

North Carolina has the second largest rural student population in the United States and a rural dropout rate on par with the national dropout rate [19]. Yet, since nearly all dropout research has been done in urban areas, little is known specifically about dropout in rural areas [4].

The motivation for this study is to compare models for predicting high school dropout for rising ninth graders in rural North Carolina. Policymakers can use these predictors to identify students who could benefit from ninth grade intervention programs as the transition to high school is a crucial time for at-risk students and ninth grade dropout prevention programs have been successful in reducing dropout rates. Researchers could use the strength and accuracy of a model in future research that uses longitudinal data to predict an educational outcome.

The research questions are:

1. What are the predictors of dropout for rising ninth graders in rural North Carolina?

2. What is the best model for predicting high school dropout?

Normally, the best predictive models minimize Type I errors of rejecting the null hypothesis when it is true. In this case, a Type I error would be saying a student is likely to drop out when they are not. Minimizing Type I errors would make an intervention program more efficient because the state would spend less money on students who didn't need the programming.

But this economic rationale does not reflect the ultimate burden that high school dropouts place on society nor does it reflect the mission that public schools should educate everyone. Therefore, in this analysis, the best predictive models are those that minimize Type II errors of failing to reject the null hypothesis when it is false. These are the cases of failing to predict that a student is likely to dropout and subsequently failing to enroll them in an intervention program. By minimizing these cases, we minimize the number of students who fall through the cracks.

1.1 Definition of Dropout and Rural

This study uses the North Carolina definition of dropout: any student who leaves school before completion of a program of studies without transferring to another elementary or secondary school [1].

In practice, this is a student who was enrolled at some time during the previous school year but who was not enrolled and did not meet reporting exclusions on day 20 of the current school year. Schools that cannot document a former student's enrollment in a US school must report that student as a dropout unless they meet one of the following exclusions:

- Students who are known to have left the country
- Students who are serving suspensions
- Students who are expelled (expelled students are counted as dropouts for federal but not state reporting)
- Students who transfer to a private school, home school, or state-approved educational program
- Students who are not enrolled on day 20 because they have serious illnesses

Students who are reported as dropouts but are not included in the dropout rate are students who leave within the first 20 days of enrollment, students incarcerated in an adult facility, and students who fail to return to school after a long-term suspension. Students who are reported as dropouts include students who leave the public schools to attend community college or who leave school to obtain a GED since it is not known whether or not they obtain a high school diploma [25]

Rural is defined using the 2010-2011 district urban-centric locale codes. Schools in areas coded as Fringe Town, Distant Town, Remote Town, Rural fringe, Rural distant, and Rural remote are considered rural

2 Literature Review

The majority of this section is paraphrased from the Literature Review section of Emily Hadley's senior public policy thesis Rural and At-Risk: Predictors

of and Methods to Address High School Dropout for Rising Ninth Graders in Rural North Carolina.

2.1 Why Does High School Dropout Matter?

In the United States, lack of a high school degree is correlated with a number of negative outcomes. Individuals who drop out are less likely to be healthy and more likely to commit a crime [41] [3]. They are less likely to find a job since 75% of jobs require more than a high school degree [26]. On average, they will make $550,000 less over the course of their lifetime and contribute an average of $139,000 less to federal and state state income taxes [3][23]. For each new high school graduate, the public health system saves $40,500 from reduced dependence on public health services while the criminal justice system saves an average of $26,600 from diminished criminal activity [23]. In 2009, North Carolina invested $11.7 million dollars in dropout prevention. The North Carolina Committee on Dropout Prevention estimated that this investment would be recovered from taxes and decreased reliance on social services if only 670 fewer students drop out[2]. This would have happened if only 20% of the students who dropped out in the cohort considered in this study had stayed in school.

2.2 Factors in High School Dropout

Students who drop out from high school often cite challenges internal and external to the school setting as reasons for dropping out. Internal challenges include strict school discipline policies and academic obstacles while external challenges include factors like out-of-school employment or having to take care of family members [35]. Since this study does not have access to data on students' reasons for dropping out, it will focus on the predictors that make a student more likely to drop out. The following factors are explored in this study:

2.3 Demographic Factors: Gender and Race

Past research has indicated that male students are less likely to graduate than female students [33] [20]. One theory for this difference is that girls display a more self-determined motivational profile than boys in high school [39]. Vallerand et al (1997) theorizes that teacher-student relationships account for some portion of this difference as female students are more likely than male students to perceive their teachers as supporting their own autonomy [39]. This lack of support combined with teachers increased propensity for critical, controlling, and punitive interactions with male students may lead to the development of a non-self-determined motivational profile that triggers undesirable consequences like dropping out of high school [39]. One limitation of the Vallerand et al (1997) study it that is was conducted in Canada, but the author suggests that it is likely generalizable to the US since similar teacher-student relationships have been observed in the US. Gender is included in this analysis to see whether status as a male student is also a significant predictor of dropout in rural areas of North Carolina.

Students from minority groups, especially Native American students, are more likely to drop out than White students [33][32]. The notable exception are Asian students who have the same or lower likelihood of dropping out when

compared to White students [12][24]. Griffin (2002) suggests that this outcome arises since Black and Hispanic students tend to show increased detachment from academics than Asian and White students [12]. Rumberger (1983) suggests that most differences in racial dropout outcomes can be explained through family characteristics like socioeconomic status [31]. Lofstrom (2007) used a Texas student sample to corroborate Rumberger's national findings and also noted the influence of community effects like the neighborhood the student lives in [24]. Race disaggregated into Asian, Black, Hispanic, Multiracial, Native American, and White is included in this analysis to see whether or not race is a significant predictor independent of the other variables included in this study.

2.4 Home Factors: Economic Disadvantage and Parent Education Level

Students who receive free or reduced price lunches are labeled as economically disadvantaged and are less likely to graduate [20][36]. The effects of a low-socioeconomic background are deeply entrenched before a student arrives for the first day of kindergarten. For example, by age four, students from high-income families have heard approximately 30 million more words than students from low-income families [6]. Students from households with low parent education levels are also less likely to graduate [31][20]. Parent education is closely connected to economic disadvantage as lower education levels often correspond to lower income levels. Yet Rumberger (1983) found evidence that education level may be more than simply an indicator of income status as it has effects that are not captured by an economic disadvantage variable alone [31]. For example, parent education levels can correspond to educational expectations for a student. Both economic disadvantage and parent education level are considered in this analysis to understand their effects as predictors of dropout.

2.5 Behavioral Factor: Days Absent

Students with a significant number of days absent in elementary or middle school [40]. The benchmark for having a significant number of days absent is usually set at 10 percent of instructional time. Silver et al (2008) found that seventh and eighth grade attendance were significant predictors of high school dropout and recognized that more days absent likely reflects lower school engagement [33]. Janosz et al (2008) finds that students with lower school engagement are more likely to drop out and suggests early identification based on disengagement factors to prevent high school dropout [16]. A count of the number of days absent in each grade from fourth grade through eighth grade is included in this analysis.

2.6 Achievement Factors: Test Scores

Student academic achievement on both math and reading exams has been positively correlated with the odds of graduation, indicating that higher achieving students are more likely to graduate [37] [33][31]. Achievement in grades below high school is important as those who fail a class in sixth or seventh grade are significantly less likely to graduate [33]. Battin-Pearson (2000) suggests that the influence of academic achievement on the likelihood of dropout is stronger than

some peer and family effects and that academic achievement should be a main focus of dropout intervention programs [7]. Reading and math end-of-grade test scores are included in this analysis.

2.7 Learning Disabilities and Limited English Proficiency

Studies appear to suggest that students with learning disabilities are more likely to drop out [21][15]. Ingrum (2006) further suggests that the effect of having a learning disability on the likelihood of dropping out is amplified by the effect of socioeconomic status [15]. Both status as learning disabled in reading and learning disabled in math are considered in this study.

While there is not a substantial amount of research on Limited English Proficiency (LEP) status in the United States and its relationship with high school dropout, LEP students in California are significantly more likely to drop out than non-LEP students [11]. Driscoll (1999) also finds that Hispanic students with lower English proficiency are more likely to drop out, even when controlling for other demographic and family background characteristics [9]. LEP status is included in this study

2.8 Retention

Students who were retained before eighth grade are significantly more likely to drop out and multiple studies have found that retention is the most significant predictor of high school dropout [17][18][30]. Students are sometimes retained in an effort to improve their academic achievement or socioemotional and behavioral adjustment, though the success of retention in affecting these outcomes is often questioned. In a metanalysis, Jimerson and Kaufman (2003) found that while some studies suggest small gains in test scores for retained students relative to similarly performing peers who were not retained, these gains typically disappear over time[18]. Roderick (1994) suggested that students who are retained are more likely to drop out because they are overage for their grade and show increasing disengagement with their school as they progress [30]. Retention before Grade 9 is considered in this analysis.

2.9 The Rural Landscape in North Carolina

In addition to identifying predictors of dropout, this study seeks to draw attention to the state of rural education in North Carolina and the state of rural education research. Currently, 49.2% of North Carolina schools are considered rural. North Carolina ranks second in the nation for both the number of rural students and the number of rural minority students. It ranks in the top ten states nationwide for percent of rural students designated as English Language Learners in 2014 (6.1%) and percent of rural students who are Title 1 eligible (23.7%). The Rural School and Community Trust releases a biannual Rural Education Priority Gauge. In the 2013-2014 report, North Carolina ranked fourth, indicating that rural North Carolina students are facing significant challenges like low achievement and high rural unemployment rates that are not being adequately addressed by state policy [19]. The body of rural education research is considerably smaller than the general body of education research. Arnold et al (2005) found only 490 research papers published between 1993 and 2001 that

specifically addressed rural education [4]. Of these, only three explored high school dropout in rural areas [4].

2.10 Unique Assets of Rural Communities

Rural America is often recognized for its small, tight-knit communities. Focus groups in rural Tennessee, Kentucky, and Alabama highlighted the importance of informal social networks in rural areas where adults in a neighborhood know and watch out for other children, ask for advice, do favors for each other, and often participate in faith-based activities [38]. Small communities reduce local government bureaucracy as every resident feels the impact of local policy and can increase inter-generational relationships [28].

Rural schools also have unique benefits. Rural schools tend to have lower student-teacher ratios compared to urban schools and parents often value the increased teacher attention [28][38]. Harde (2003) applied self-determination theory to education and suggested that the effect of a teacher's support on a student's motivation was noticeably stronger for rural students, potentially indicating that rural students' academic motivation is closely related to the quality of their teachers' motivating styles [13].

2.11 Unique Challenges Faced by Rural Communities and Schools

A major challenge in many rural areas is unemployment. The North Carolina rural unemployment rate is around 8.6%, the seventh highest in the nation for rural areas [19]. Mining and farming jobs have been replaced by low-wage jobs in retail and service industries, contributing to the economic disadvantage of these areas. In focus groups in rural Tennessee, Kentucky, and Alabama, the majority of participants said that if they had a magic wand, they would create more jobs with higher wages as participants believe that a higher income would improve the lives of their children and the community [38].

These low-wage environments contribute to increased poverty. The rural child poverty rate in North Carolina is higher than the urban child poverty rate in 19 states [34]. North Carolina ranks eighth out of all states for the number of students who are Title 1 eligible with 23.7% of students meeting the qualifying standards [19]. Families in rural areas often have high property taxes, sometimes up to half of a family's income in very poor areas [28]. Yet rural areas still have a smaller tax base when compared to urban areas [28].

Rural schools face a number of challenges due to limited funding and small school size. First, rural schools tend to have lower teacher salaries and subsequently struggle to recruit high qualified teachers and teachers with advanced degrees [10] [28]. Rural schools also struggle to offer advanced courses like AP classes and are forced to limit counseling and psychological services [10]. Students travel significant distances to get to school and many may not be able to participate in after school activities since they must take the only available bus home right after school ends. There also are not as many extracurricular opportunities in rural areas due to the small school size, contributing to the discovery that the highest level of drug use for youth ages 12 to 17 is in rural communities [22].

Finally, rural communities tend to have less exposure to diversity and higher education. Though North Carolina has the ninth largest rural minority population and ranks eighth for number of ELL students, many rural areas are not exposed to racial or cultural diversity as they have been a majority white for generations. Adults in rural areas are more likely to have a high school degree than adults in non-rural areas, but rural adults are considerably less likely to have a college degree [34]. Parents in the Tennessee, Kentucky, and Alabama focus groups recognized that the lack of higher education in rural areas may be contributing to poor early childhood home environments relative to peers in more educated households [38]

2.12 Existing Education Models

Other studies that have looked at education data have generally used frequentist models for their analysis. The most common is crosstabular analysis with some frequentist inference for means [40] [26]. The next step up in complexity is to analyze the data using a logistic or probit regression [36] [20]. The most complex models are hierarchical generalized linear models that consider both school and student level effects [37]. I intend to build and compare both logistic regression models and HGLM for this data to Bayesian hierarchical regression models.

3 Data Overview

This section is paraphrased from the Literature Review section of Emily Hadley's senior public policy thesis Rural and At-Risk: Predictors of and Methods to Address High School Dropout for Rising Ninth Graders in Rural North Carolina.

The data for this analysis was provided by the North Carolina Department of Public Instruction. The North Carolina Education Research Data Center (NCERDC) created this dataset by concatenating the data across school years to create one file per student. NCERDC also provided files on each school attended by a student in the cohort. School indicators were used for grouping in the hierarchical model.

3.1 Student Data

The data consists of the 53,996 students initially enrolled in third grade at any North Carolina school in 2002-2003, who remained enrolled in a North Carolina school through the expected eighth grade year of 2007-2008, and who were enrolled in a school classified as rural in their expected eighth grade year. Not all students in the cohort were actually in eighth grade in the 2007-2008 school year as some students were retained.

Data was collected on this cohort from third grade through expected graduation in 2012. Students who moved to North Carolina after the 2002-2003 school year were not included as they were not added to the cohort. Students who left the cohort before their expected eighth grade year were removed from the dataset since the goal of this analysis is to predict the likelihood of dropout for rising ninth graders. Approximately 3.9% of the cohort (2068 students) dropped out. This is slightly above the national 2011-2012 dropout percentage of 3.3%.

Table 1: Percentage and Count of Students By Race and Gender

Gender	White	Black	Asian	Hispanic	Native American	Multiracial	Total
Male	31.1 (16786)	14.0 (7563)	0.6 (311)	3.1 (1673)	1.3 (675)	0.9 (512)	50.9 (27520)
Female	30.1 (16254)	13.3 (7197)	0.5 (281)	3.0 (1634)	1.1 (600)	0.9 (510)	49.1 (26476)
Total	61.2 (33040)	27.3 (14760)	1.1 (592)	6.1 (3307)	2.4 (1275)	1.9 (1022)	100 (53996)

The student-level dataset initially contained 148 variables. Predictors collected in high school, such as End of Course test scores and days absent in ninth, tenth, eleventh, or twelfth grade, were removed as the goal of this analysis is to predict dropout based on data available through eighth grade. The final student-level dataset was reduced to 34 variables. A full list of variables is in Appendix A.

3.2 Demographic Variables: Gender and Race

Table 1 illustrates the racial and gender distribution of the students in the cohort. The largest racial group in the cohort is white students followed by African American students, Hispanic students, and Native American students. Multiracial and Asian students compose an especially small proportion of the cohort. There are slightly more male students than female students.

3.3 Dropout, Economic Disadvantage, LEP, and Retention

Table 2 further disaggregates the gender and race data by other risk factors. This table illustrates that male and female students in the cohort have similar rates of being identified as economically disadvantaged (ED) or Limited English Proficiency (LEP). Male students are more likely to drop out, more likely to be retained, and more likely to have a learning disability in reading (LDR) or a learning disability in math (LDM).

Table 2 also reveals racial disparities among these risk factors. Native American males and females have remarkably high rates of dropout. African American and Hispanic students have high rates of economic disadvantage. More than 90% of both male and female Hispanic students qualify for free or reduced price lunch. Asian and Hispanic males and females have high rates of LEP status which is expected since these are two of the largest immigrant groups in America. One particularly notable statistic is that more than 20% of the Hispanic, Native American, and Multiracial male students were retained before eighth grade. A different retention trend was seen for female students as Black female students were most likely to be retained.

Table 2: Percentage and Count of Students by Gender, Race, and Risk Factors

Males	Drop Out	ED	LEP	Retained	LDR	LDM
White	4.1	44.9	0.4	6.1	14.0	9.8
	(700)	(7548)	(64)	(1031)	(2354)	(1646)
Black	5.4	87.9	0.3	14.0	18.1	14.5
	(412)	(6646)	(19)	(1061)	(1371)	(1100)
Asian	1.9	67.9	58.5	3.9	6.8	4.8
	(6)	(211)	(182)	(12)	(21)	(15)
Hispanic	4.7	94.5	88.2	28.9	16.2	12.2
	(79)	(1581)	(1476)	(484)	(271)	(204)
Native American	8.0	88.6	0.0	26.1	20.7	15.7
	(54)	(598)	(0)	(176)	(140)	(106)
Multiracial	4.5	75.2	3.5	24.2	16.4	12.7
	(23)	(385)	(18)	(124)	(84)	(65)
All Males	4.6	61.7	6.4	8.8	15.4	11.4
	(1274)	(16969)	(1759)	(2400)	(4241)	(3136)
Females	Drop Out	ED	LEP	Retained	LDR	LDM
White	2.8	45.3	0.3	3.8	7.4	5.4
	(455)	(7355)	(54)	(611)	(1210)	(873)
Black	3.4	88.4	0.2	8.0	9.3	(7.9)
	(248)	(6362)	(16)	(572)	(670)	(569)
Asian	.4	61.6	55.5	1.4	5.7	4.3
	(1)	(173)	(156)	(4)	(16)	(12)
Hispanic	2.9	93.3	87.6	6.1	9.4	7.1
	(48)	(1523)	(1431)	(100)	(154)	(116)
Native American	5.2	87.0	0.8	6.0	10.2	7.8
	(31)	(522)	(5)	(36)	(61)	(47)
Multiracial	2.2	72.7	3.7	3.5	8.8	5.7
	(11)	(371)	(19)	(18)	(45)	(29)
All Females	2.9	61.6	6.3	5.1	8.1	6.2
	(794)	(16306)	(1681)	(1341)	(2156)	(1646)

Percentage and count (in parentheses) of each gender disaggregated by race and risk factors. Example interpretation is that 4.1% of White, Male students drop out. ED = Economically Disadvantaged; LEP = Limited English Proficiency; LDR = Learning Disability in Reading; LDM = Learning Disability in Math

Table 3: Mean and SD of Days Absent by Dropout Status and Grade

Days Absent	Dropout		Non-Dropout	
Grade	Mean	SD	Mean	SD
Fourth	8.6	7.6	6.2	5.8
Fifth	8.9	7.4	6.4	6.2
Sixth	11.1	9.8	7.2	7.1
Seventh	13.1	11.1	8.1	8.2
Eighth	14.9	12.7	8.5	9.1

SD = Standard Deviation

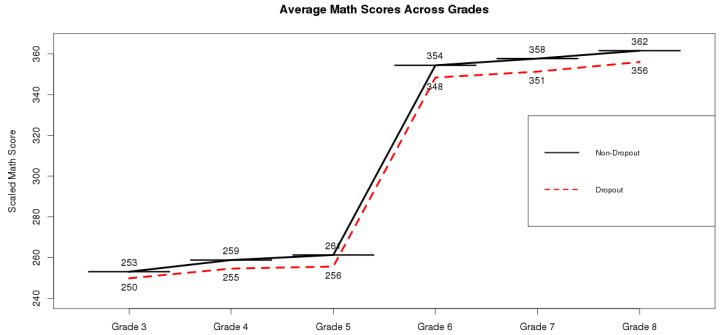

Figure 1: This figure illustrates the difference between the average math test scores across grades for students who do not drop out and the average math test scores across grades for students who do drop out.

3.4 Days Absent

Table 3 shows that in fourth grade, students who go on to drop out already have more absences on average than students who do not go on to drop out.[1] This trend widens over time as students who go on to drop out see a 6.3 day increase in the average number of days absent while the students who do not go on to drop out see a 2.3 day increase in the average number of days absent.

3.5 Test Score Data

Figure 1 and Figure 2 show that, starting in third grade, students who do not go on to drop out have higher average test scores in both reading and math. The gap in reading scores remains around five points while the gap in math scores increases from around three points to between six and seven points. Since the test scores are scaled, they are normally distributed around the mean. Due to this normalcy, Welch's Two Sample T-Test was used to confirm that the average test scores in both math and reading for students who went on to dropout were lower than the average test scores for students who did not go on to drop out at a statistically significant level.

3.6 School Data

Data was also collected for the 1226 rural schools attended by the students in this cohort. The strongest correlation between dropout and a school-level variable was a small 0.05 correlation between dropout and the average percent of students proficient in math in a particular school. This suggests that student-level variables are better predictors of dropout than school-level variables. Since school factors were incorporated into the final model by using multilevel analysis which intrinsically controls for school-level effects, individual school variables were not included in the model as they would be redundant.

[1] Third grade absence data was not available in this dataset

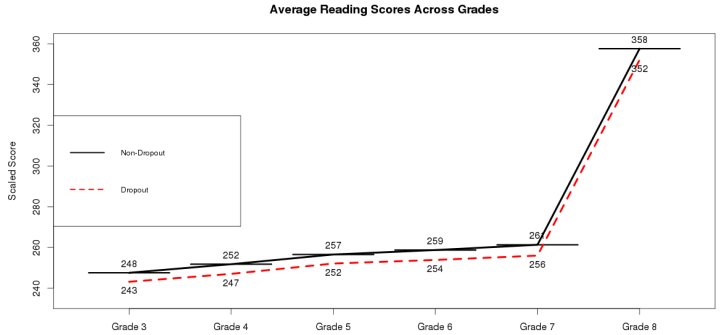

Figure 2: This figure illustrates the difference between the average reading test scores across grades for students who do not drop out and the average reading test scores across grades for students who do drop out.

3.7 Missing Data

While the outcome variable of dropout status is recorded for all students, some predictor values were missing. These included items like test score data for students who were retained, learning disability indicators, and LEP status. Rather than eliminate cases with missing data, multivariate imputation through chained equations in the **MICE** package in **R** was used to impute all missing values.[2]

3.8 The Limitation of "Ever Identified"

One limitation of this dataset is that the following variables are based on whether students were *ever identified* as having these characteristics. Since the initial dataset went from third grade through high school, it is possible that the students in the cohort did not have the following characteristics until high school:

- Economic Disadvantage
- Learning Disability in Reading
- Learning Disability in Math
- Limited English Proficiency

This discrepancy threatens the analysis because the models in the subsequent chapters are attempting to predict dropout based on data from third through eighth grade and it is impossible to know when students were assigned these labels.

Yet, there is reason to believe that late labeling in high school for these four variables is unlikely. Late labeling as economically disadvantaged is unlikely as students in high school are less likely to sign up for free/reduced price lunch. If students were ever going to be labeled as economically disadvantaged, it is more

[2]The **MICE** package uses multivariate imputation through chained equations which is an iterative process that imputes missing variables based on the other observations in the dataset. A detailed description is available in Azur et al (2012) [5].

likely that they would first receive the label in elementary school [29]. It is also unlikely that a student would not have been identified as Learning Disabled in Reading, Learning Disabled in Math, or Limited English Proficiency until high school since all students in this study were part of the 2002-2003 third-grade cohort and it is unlikely that these characteristics would not have been noticed until ninth grade or after. This study therefore assumes that these labels were assigned before high school while acknowledging that this assumption is still a limitation of the analysis.

4 Methodology

Due to computational issues, these models were built using 9278 students from 50 randomly-selected rural schools. Before modeling, k-means clustering was used to cluster variables tracked across time. Since the goals of this analysis include both prediction and comparison of models, a wide variety of methods have been considered. The first two regression methods are common in social sciences: multivariate logistic regression and multilevel, multivariate logistic regression. The third method uses principle component analysis in a principle component regression. The final method utilizes Bayesian regression to update priors with data.

4.1 K-Means Clustering

K-means clustering is an effective way of summarizing data across time to minimize the effects of multicollinearity between years. In k-means clustering, an iterative algorithm is used to assign each data point to the cluster with the nearest mean. The number of clusters is fixed at the number where adding another cluster doesn't lead to better modeling of the data.[3] In this study, k-means clustering was utilized to summarize data for math scores from grades 3-8, reading scores from grades 3-8, and absence data from grades 4-8. Each student was placed in one of six reading and math score clusters. Each student was also placed in one of five absence clusters.

Table 4 shows the six score clusters that each student is assigned to. For example, a student who consistently performed 1.6 standard deviations above the average test score would, through the k-means algorithm, be placed in the "Considerably Above Average" cluster group. Table 4 also shows the mean for each cluster across all six years of data collection. Distinct differences are clear between each cluster, indicating that individual students often follow a pattern of high, average, or low performance. The clusters with the most students are those that are closest to the average while the clusters with the smallest numbers of students are the furthest from the average. This distribution of students is expected since the test score data for all grades is distributed normally.

Table 5 shows the mean number of days absent for each cluster across five years of data collection. While differences between clusters are not as defined as in the test score clusters, each cluster still exhibits a pattern that is described by the cluster name. The Increasing Absence, Moderate Absence, and Severe Absence clusters are all notable as they each have at least one grade where the

[3] Hartigan and Wong (1979) give a detailed description of a K-Means clustering algorithm [14].

Table 4: SD of Test Scores in Reading and Math Clusters Across Grades

Reading Cluster — Grade	3rd	4th	5th	6th	7th	8th	N
Considerably Above Average	1.36	1.41	1.42	1.41	1.39	1.50	4616
Above Average	0.74	0.77	0.76	0.77	0.77	0.78	8116
Slightly Above Average	0.21	0.20	0.21	0.22	0.22	0.18	11269
Slightly Below Average	-0.31	-0.34	-0.32	-0.33	-0.35	-0.42	12544
Below Average	-0.95	-0.96	-0.96	-0.97	-0.97	-1.01	11338
Considerably Below Average	-1.76	-1.78	-1.83	-1.85	-1.78	-1.57	6113
Math Cluster — Grade	3rd	4th	5th	6th	7th	8th	N
Considerably Above Average	1.56	1.59	1.72	1.65	1.70	1.75	5490
Above Average	0.85	0.90	0.94	0.97	0.97	0.94	10445
Slightly Above Average	0.33	0.35	0.33	0.37	0.34	0.25	12374
Slightly Below Average	-0.17	-0.20	-0.26	-0.27	-0.27	-0.33	11798
Below Average	-0.72	-0.78	-0.85	-0.95	-0.94	-0.80	9131
Considerably Below Average	-1.71	-1.68	-1.55	-1.43	-1.40	-1.33	4758

N=Number of students in each cluster

Table 5: Number of Days Absent in Each Grade by Absence Cluster

Attendance — Grade	4th	5th	6th	7th	8th	N
Low Absence	3.14	2.93	3.13	3.45	3.70	24837
Mild Absence	7.87	7.92	8.51	8.67	8.09	17368
Increasing Absence	7.15	7.54	9.99	14.62	20.0	5384
Moderate Absence	15.66	17.13	19.97	19.25	16.09	5152
Severe Absence	15.15	17.65	24.53	37.37	46.30	1255

N=Number of students in each cluster

average number of days absent is more than 10% of the school year (generally 18 days). The cluster with the most students is the cluster with the lowest number of average absences and the cluster with the fewest students is the cluster with the highest number of average absences. This distribution of students is expected since absence data exhibits a right skew where students are far more likely to have fewer absences.

4.2 Logistic Regression

Logistic regression is often used to model situations with binary outcomes. In this study, the binary outcome is 'dropout' or 'non-dropout'.

The odds of a student being a dropout are defined as the probability that a particular student is a dropout π_i divided by the probability that a particular student is not a dropout $1-\pi_i$. Equation 1 shows how logistic regression predicts the natural log of these odds in terms of an intercept B_0, a series of predictors X_n, and their respective coefficients B_n.

$$ln(\frac{\pi_i}{1-\pi_i}) = B_0 + B_n X_n \tag{1}$$

The coefficients B_n for the predictors X_n represent the change in the logit outcome for a unit increase in the predictor.

4.3 Multilevel Logistic Model

Multilevel modeling is used to control for school level effects since many students in this study attend the same schools. Multilevel modeling can correct biases in some parameter and standard error estimates as it adjusts for correlation effects.

The multilevel equation is:

$$y_{ij} = B_0 + B_n x_n + u_j \tag{2}$$

In Equation 4.5, y_{ij} is the log odds of dropping out for the ith student at the jth school. Like in the logistic model, B_0 is the intercept, X_n is the explanatory variable for the ith student at the jth school, and B_n is the accompanying coefficient. u_j is a random effect accounting for variation at the school level.

4.4 Principle Component Analysis and Principle Component Regression

Principle Component Analysis (PCA) and the associated Principle Component Regression (PCR) are ways to reduce the dimensions of a given dataset. PCA attempts to produce a small set of independent principle components from a larger set of original values. Equation 3 shows how principle component values Y are defined as linear combinations of other variables X.

$$Y_k = C_{k1}X_1 + C_{k2}X_2 + ... + C_{km}X_m \tag{3}$$

A scree plot can be used to determine how many principle components to retain as one usually takes all components before the line plateaus. The outcome variable, in this case dropout, is then regressed on the selected components. The model can then be used to predict other cases.

4.5 Bayesian Hierarchical Regression

A Bayesian model was used to borrow information across schools which is useful as some schools have small numbers of dropouts. The model is:

$$log(\frac{\pi_{ij}}{1-\pi_{ij}}) = u_j + B_n x_n \tag{4}$$

Here, the left side of the equation is the log odds that student i in school j drops out. The prior distributions for random intercepts u_j and the coefficients of B_n were drawn as follows[4]:

$$\begin{aligned} B_n \ & N(0, \sigma^2) \\ u_i \ & N(0, \tau^-1) \\ \tau^-1 \ & Cauchy_[0, inf](0, 1) \\ \sigma^2 \ & Cauchy_[0, inf](0, 1) \end{aligned}$$

[4]The following was inspired by similar work and topics covered in Statistics 601: Bayesian Statistics, Spring 2014

A half-cauchy prior was used instead of the more commonly-used inverse prior as it has been found to be weakly informative and displays better properties in Polson and Scott (2012) and Britain et al (2014) [27] [8].

5 Results

These models all attempt to use the same variables to try to allow for comparison. Parent Education level was not included as it has many levels and led to computational concerns in Bayesian analysis. Race coefficients were compared to the baseline of White. Due to computational challenges, clusters were treated as numeric values rather than factors. The higher the cluster number, the higher the student performed on reading or math tests, or the higher the number of absences.

5.1 Results from Logistic Regression

Table 6 shows the results from the logistic regression. The odds of a male student dropping out were 1.48 times larger than the odds of a female student dropping out. The odds of an American Indian student dropping out are 3.78 times larger than the odds of a White student dropping out. The odds of other races dropping out are not significantly different than the odds of White students dropping out. The odds of a student ever identified as economically disadvantaged dropping out are 2.75 times larger than the odds of a non-economically disadvantaged student dropping out. The odds of a retained student dropping out are 1.62 times larger than the odds of a student who was not retained dropping out. For every increase in Days Absent Cluster (for example, moving from Mild to Moderate Absences), the odds of dropout are 1.21 times larger. The odds of an LEP student dropping out are 0.34 times smaller than the odds of a non-LEP student. For every increase in math cluster (for example, moving from Slightly Below Average to Slightly Above Average), the odds of dropout are 0.75 times smaller.

Yet, this model does not have great fit as the Confusion Matrix in Tableřefconfusionlog shows that the model only correctly predicts 71% of students who do not drop out in a test set and 65% of students who do drop out in a test set. Figure **??** reinforces that the model does not fit as it shows that the predicted likelihoods of dropout for students who actually go on to drop out in the test set are not substantially different than the predicted likelihoods of dropout for students who do not go on to drop out. Finally, the Hosmer-Lemeshow test returned a significant p-value, indicating the model had poor fit. Interactions were not considered to allow for comparison between models.

5.2 Results from Multilevel Logistic Regression

Table 8 shows the results from the multilevel logistic regression that incorporated school-level random effects.[5] Male students have odds of dropping out 1.49 times larger than the odds for female students. American Indian students are again the only race significantly different than White and they have odds of

[5]This is NOT the same model used in Emily Hadley's senior public policy thesis as that model included variables that are not in this analysis and a larger dataset

Table 6: Results from Logistic Regression

	Dependent variable:
	dropout
Intercept	-3.898^{***}
	(0.312)
Male	0.394^{***}
	(0.134)
Asian	-0.692
	(1.059)
American Indian	1.327^{**}
	(0.579)
Black	-0.118
	(0.150)
Hispanic	0.708
	(0.504)
Multiracial	-0.865
	(0.724)
Ever ED	1.101^{***}
	(0.205)
Ever LEP	-1.065^{*}
	(0.552)
Retained	0.489^{***}
	(0.177)
Days Absent Cluster	0.188^{***}
	(0.053)
Math Cluster	-0.293^{***}
	(0.057)
Observations	7,422
Log Likelihood	$-1{,}015.808$
Akaike Inf. Crit.	2,055.616
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 7: Confusion Matrix for Logistic Regression

	Students Not Predicted to Drop Out	Students Predicted to Drop Out
Students Who Do Not Drop Out	5058	2107
Students Who Do Drop Out	168	89

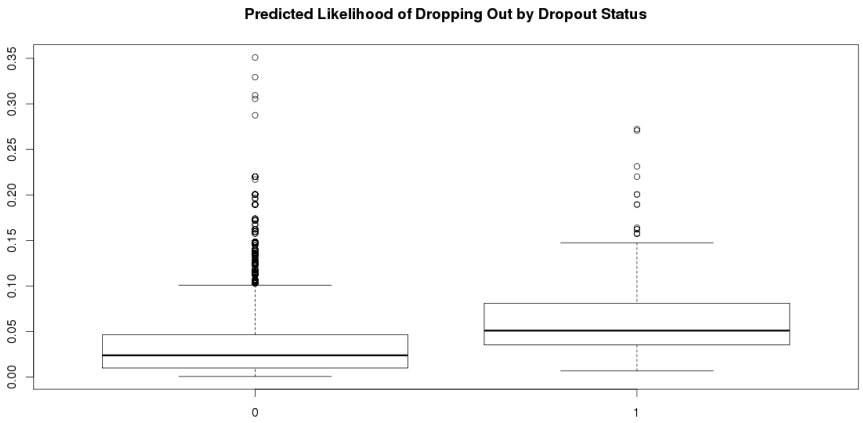

Figure 3: Boxplot of Predicted Likelihood of Dropout versus Actual Dropout in Logistic Regression

dropping out 3.90 times larger than the odds for White students. Economically disadvantaged students have odds of dropping out 3.03 times higher than the odds for non-economically disadvantaged students. The odds of a retained student dropping out are 1.61 times larger than the odds of a student who was not retained dropping out. For every increase in Days Absent Cluster (for example, moving from Mild to Moderate Absences), the odds of dropout are 1.21 times larger. The odds of an LEP student dropping out are 0.34 times smaller than the odds of a non-LEP student. For every increase in math cluster (for example, moving from Slightly Below Average to Slightly Above Average), the odds of dropout are 0.75 times smaller.

Yet, this model also does not have great fit as the Confusion Matrix in Tableřefconfusionmulti shows that the model only correctly predicts 71% of students who do not drop out in a test set and 68% of students who do drop out in a test set. Figure **??** reinforces that the model does not fit as it shows that the predicted likelihoods of dropout for students who actually go on to drop out in the test set are not substantially different than the predicted likelihoods of dropout for students who do not go on to drop out. Finally, the Hosmer-Lemeshow test returned a significant p-value, indicating the model had poor fit. Interactions were not considered to allow for comparison between models.

5.3 Results from Principal Component Analysis and Principle Component Regression

Principle Component Analysis (PC) was completed for numeric student-level variables in a randomly-selected training dataset. All numeric variables were standardized before completing PCA.

Based on Figure 5, four components were used in the PCA. Table 10 shows the weight of each component in the Principle Component Regression (PCR). PCR places weight on the individual components rather than the individual variables, so it is not possible to understand the significance of each individual

Table 8: Results from Multilevel Logistic Regression

	Dependent variable:
	dropout
School Random Effects	Variance = 0.05
Intercept	-3.921^{***}
	(0.317)
Male	0.399^{***}
	(0.134)
Asian	-0.664
	(1.061)
American Indian	1.360^{**}
	(0.584)
Black	-0.127
	(0.156)
Hispanic	0.722
	(0.506)
Multiracial	-0.883
	(0.725)
Ever LEP	-1.073^{*}
	(0.554)
Days Absent Cluster	0.187^{***}
	(0.054)
Ever ED	1.105^{***}
	(0.206)
Retained	0.479^{***}
	(0.179)
Math Cluster	-0.293^{***}
	(0.058)
Observations	7,422
Log Likelihood	$-1{,}015.187$
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 9: Confusion Matrix for Multilevel Logistic Regression

	Students Not Predicted to Drop Out	Students Predicted to Drop Out
Students Who Do Not Drop Out	5068	2097
Students Who Do Drop Out	175	82

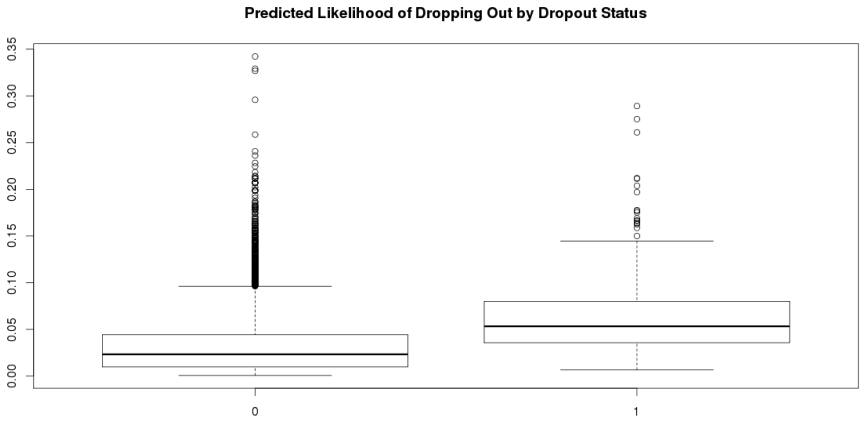

Figure 4: Boxplot of Predicted Likelihood of Dropout versus Actual Dropout in Multilevel Logistic Regression

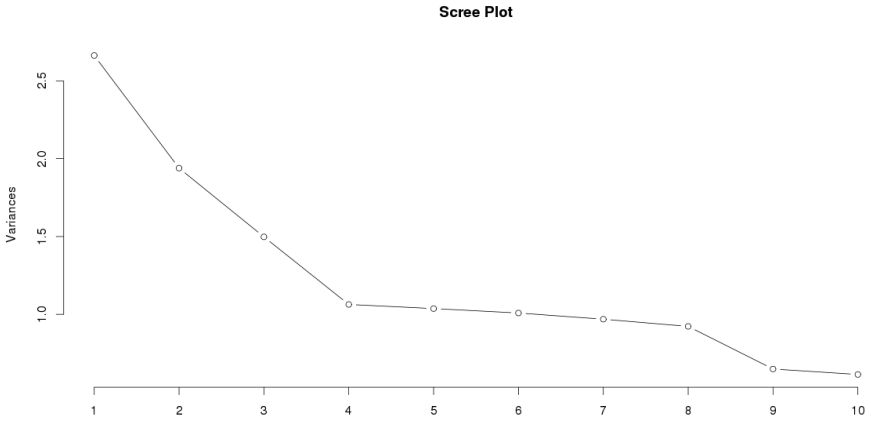

Figure 5: Scree Plot for PCA

variable in this model on the outcome of dropout.

Table 10: Components of PCA

	Dependent variable:
	dropout
Constant	-6.279^{***}
	(0.227)
Component 1	0.718^{***}
	(0.061)
Component 2	-0.817^{***}
	(0.073)
Component 3	0.443^{***}
	(0.065)
Component 4	-2.382^{***}
	(0.116)
Observations	7,422
Log Likelihood	-458.111
Akaike Inf. Crit.	926.223
Note:	*p<0.1; **p<0.05; ***p<0.01

Yet PCA appears to be a good predictive model for high school dropout in rural North Carolina. Figure 6 illustrates that the PCA model does a decent job of assigning higher predictor values to students who go on to drop out. This is further demonstrated in the confusion matrix in 11 as the 98% of dropouts are correctly predicted as dropouts and 92% of non-dropouts are correctly predicted as non-dropouts.

5.4 Results from Bayesian Analysis

Computation was done using JAGs. Computation issues were a serious concern for this model as the R host would crash with too many variables or too many iterations. Even with only 250 iterations from two chains and 25 discarded, the model took 8 hours to run. Since so few iterations could be used, I do not have confidence in the convergence of this model. This is verified in trace plots and density plots for each variable that reveal poor convergence. If I had more computing power, I would run this model with significantly more iterations.

Table 12 shows the results of the Bayseian model. Here, Asian, Black, and Multiracial students are less likely to drop out than White students with odds 0.46, 0.47, and 0.41 times smaller respectively. American Indian and Hispanic

Table 11: Confusion Matrix for PCA

	Students Not Predicted to Drop Out	Students Predicted to Drop Out
Students Who Do Not Drop Out	6624	541
Students Who Do Drop Out	6	251

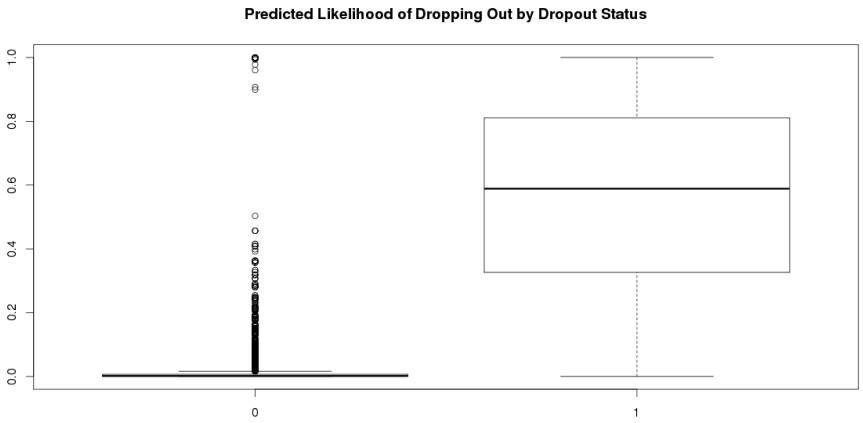

Figure 6: Boxplot of Predicted Likelihood of Dropout versus Actual Dropout in PCA

Table 12: Results from Bayesian Model

Variable	Coef	SD	Exponentiate
Male	-0.02	0.08	0.98
American Indian	0.52	0.28	1.68
Asian	-0.77**	0.41	0.46
Black	-0.75***	0.14	0.47
Hisp	-0.01	0.41	0.99
Mult	-0.89*	0.49	0.41
Ever ED	0.07	0.16	1.07
Ever LEP	-0.64	0.37	0.53
Days Absent Cluster	-0.11*	0.05	0.90
Math Cluster	-0.80***	0.07	0.45
Retained	0.01	0.12	1.01

students did not have odds that were significantly different than those of White students. For each increase in Days Absent Cluster (for example, moving from Mild Absence to Moderate Absence), students had 0.9 times smaller odds of dropping out and for each increase in Math Cluster, students had 0.45 times smaller odds of dropping out. Economic disadvantage, LEP status, and retention were not significant predictors.

6 Discussion

This is not a causal study so the results should not be used to imply causality. For example, policymakers should not use these results to say that low math achievement causes dropout. Policymakers could consider using these results to identify the most important predictors of dropout even when controlling for other predictors and thereby identify rising ninth graders in rural areas who are

at-risk of dropping out.

However, the results in this analysis reveal the challenge of finding an accurate, interpretable model when working with multilevel, longitudinal education data. The logistic regression, the multilevel logistic regression, and the Bayes model all have coefficients that can be interpreted.

The logistic regression and the multilevel regression suggest that policymakers should target a ninth grade intervention program at Male students, American Indian students, Economically Disadvantaged students, Retained students, students with many absences, and students with low math scores. Interestingly, these models suggest that policymakers should not target LEP students for an intervention program. Perhaps this is because these students are being adequately served by the small classrooms of rural America or perhaps this is capturing some other effect. The multilevel model was slightly better at predicting the likelihood of high school dropout, but both the logistic regression and multilevel regression models had poor fit, perhaps because there were so few dropouts in the random sample. Policymakers should be cautious before using these results to inform policy[6].

The PCA is the most accurate model considered in this analysis for predicting dropout. However, it does not allow for analysis of the most significant variables of high school dropout. If policymakers are strictly interested in predicting the likelihood of high school dropout and not interested in the actual risk factors themselves, policymakers should consider incorporating PCA into the process of identifying at-risk students.

The Bayes model has the potential to accurately reflect the data in an interpretable way. Unfortunately, the final model is this analysis did not have good convergence. This was likely the reason the results were so different from those of the logistic regression. For example, no other model I tested said that economic disadvantage and retention were not significant predictors of dropout. A Bayes model should not be used for predicting high school dropout unless the computational power issues can be solved.

7 Conclusion

Policymakers could use multilevel logistic regression models to help identify rural students who are at-risk of dropping out, but should consider using the full sample and incorporating more variables or interactions into the model to get better model fit. If policymakers are more interested in prediction than interpretation, PCA is an accurate method though it cannot be used to explore which predictors are most important. Finally, Bayes modelling shows promise in educational research but struggles with issues of computational power.

[6]I have more confidence in the model I used in my senior public policy thesis which is different from this model

References

[1] Policy Regarding dropout prevention and students at-risk, 2004.

[2] Model Dropout Prevention Programs, 2011.

[3] Solving the Graduation Crisis: Identifying and Using School Feeder Patterns in your Community, 2013.

[4] Michael Arnold, John Newman, Barbara Gaddy, and Ceri Dean. A Look at the Condition of Rural Education Research: Setting a Direction for Future Research. *Journal of Research in Rural Education*, 20(6), 2005.

[5] Melissa Azur, Elizabeth Stuart, Constantine Frangakis, and Philip Leaf. Multiple imputations by Chained Equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research*, 20(1):40–49, March 2012.

[6] Daphna Bassok and Loeb Susanna. Early Childhood and the Achievement Gap. In Ladd Helen and Goertz Margaret, editors, *Handbook of Research in Education Finance and Policy*, pages 510–527. Routledge, New York, 2 edition, 2015.

[7] Sara Battin-Pearson. Predictors of early high school dropout: A test of five theories. *Journal of educational psychology*, 92(3):568–582, 2000.

[8] Travis Britain, Emily Hadley, Meghan Scanlon, and Heather Shapiro. Helmet Laws: A No Brainer. 2014.

[9] Anne Driscoll. Risk of High School Dropout among Immigrant and Native Hispanic Youth. *International Migration Review*, 33(4):Winter 1999, January 857.

[10] Rebecca Droessler Mersch. *Student Academic Achievement in Rural vs. Non-Rural High Schools in Wisconsin.* Dissertation, Edgewood College, 2012.

[11] Patricia Gandara. Review on Research on the Instruction of Limited English Proficient Students: A Report to the California Legislature. Technical report, 1997.

[12] Bryan Griffin. Academic Disidentification, Race, and High School Dropouts. *The High School Journal*, 85(4):71–81, May 2002.

[13] Patricia Harde and John Marshall Reeve. A motivational model of rural students' intention to persist in, versus drop out of, high school. *Journal of Educational Psychology*, 95(2):347–356, June 2003.

[14] JA Hartigan and MA Wong. A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, 28(1):100–108, 1979.

[15] Adrienne Ingrum. High School Dropout Determinants: The Effect of Poverty and Learning Disabilities. *The Park Place Economist*, 14(1):73–79, 2006.

[16] Michel Janosz, Isabelle Archambault, Julien Morizot, and Linda Pagani. School Engagement Trajectories and Their Differential Predictive Relations to Dropout. *Journal of Social Issues*, 64(1):21–40, 2008.

[17] Shane Jimerson, Gabrielle Anderson, and Angela Whipple. Winning the Battle and Losing the War: Examining the Relation Between Grade Retention and Dropping Out of High School. *Psychology in the Schools*, 39(4), 2002.

[18] Shane Jimerson and Amber Kaufman. Reading, Writing, and Retention: A Primer on Grade Retention Research. *The Reading Teacher*, 56(7):622–635, April 2003.

[19] Jerry Johnson, Daniel Showalter, Robert Klein, and Christine Lester. Why Rural Matters 2013-14. Technical report, May 2014.

[20] J.L. Jordan, G. Kostandini, and E. Mykerezi. Rural and urban high school dropout rates: Are they different? *Journal of Research in Rural Education*, 21(12):1–21, 2012.

[21] Larry Kortering and Patricia Braziel. School Dropout among Youth with and without Learning Disabilities. *Career Development and Transition for Exceptional Individuals*, 21(1):61–74, April 1998.

[22] David Lambert, John Gale, and David Hartley. Substance Abuse by Youth and Young Adults in Rural America. *The Journal of Rural Health*, 24(3):221–228, 2008.

[23] Henry Levin, Clive Belfield, Peter Muennig, and Cecilia Rouse. The Costs and Benefits of an Excellent Education for All of America's Children, 2006.

[24] Magnus Lofstrom. Why are Hispanic and African-American Dropout Rates So High? In *IZA Discussion Paper*, December 2007.

[25] J. Owen, J. Rosch, C Muschkin, J. Alexander, and C. Wyant. Dropout prevention: Strategies for improving high school graduation rates. *Center for Child and Family Policy*, 2008.

[26] E. Pascarella and P. Terenzini. Predicting freshman persistence and voluntary dropout decisions from a theoretical model. *The Journal of Higher Education*, 51(1):60–75, 1980.

[27] Nicholas Polson and James Scott. On the Half-Cauchy Prior for a Global Scale Parameter. *Bayesian Analysis*, 7(4):887–902, 2012.

[28] Allan Porowski and Caitlin Howley. Dropout Prevention: Challenges and Opportunities in Rural Settings, 2013.

[29] Katherine Ralston, Constance Newman, Annette Clauson, Joanne Guthrie, and Jean Buzby. The National School Lunch Program: Background, Trends, and Issues. Technical Report Economic Research Report 61, USDA, July 2008.

[30] Melissa Roderick. Grade Retention and School Dropout: Investigating the Association. *American Educational Research Journal*, 31(4):729–759, 1994.

[31] Russell Rumberger. Dropping out of High School: The Influence of Race, Sex, and Family Background. *American Educational Research Journal*, 20(2):199–220, 1983.

[32] Kelsey Sheehy. Graduation Rates Dropping Among Native American Students. *US News*, June 2013.

[33] D. Silver, M. Saunders, and E. Zarate. What Factors Predict High School Graduation in the Los Angeles Unified School District, 2008.

[34] Jay Smink and Mary Reimer. Rural School Dropout Issues: Implications for Dropout Prevention Strategies and Programs.

[35] E. Stearns and E. Glennie. When and Why Dropouts Leave High School. 2006.

[36] B.R. Subdei and B. Johnson. Predicting High School Graduation and Dropout Using a Hierarchical Generalized Linear Model Approach. 2007.

[37] R.S. Subedi and M. Howard. Predicting high school graduation and dropout for at-risk students: A multilevel approach to measure school effectiveness. *Advances in Education*, 2(1):11–17, 2013.

[38] Linda Tilly, Apreill Curtis Hartsfield, Lisa Parrish, Debra Miller, Valerie Salley, Linda O'Neal, Pam Brown, and Edwina Chappell. The Rural South: Listening to Families in Alabama, Kentucky, and Tennessee, 2004.

[39] Robert Vallerand, Michelle Fortier, and Frederic Guay. Self-determination and persistence in a real-life setting: Toward a motivational model of high school droput. *Journal of Personality and Social Pscyhology*, 72(5):1161–1176, May 1997.

[40] T.C. West. Just the Right Mix: Identifying Potential Dropouts in Montgomery County Public Schools Using an Early Warning Indicators Approach, 2013.

[41] O. Yeboah, P.E. Faulkner, and G. Appiah-Danquah. North Carolina High School Dropout Rates: An Econometric Analysis. 2010.