THE UNIVERSITY OF CHICAGO

Predicting Dropouts in Montgomery County Public Schools

June 8, 2015

PRIMARY INVESTIGATORS:

JOSHUA MAUSOLF

BRIDGIT DONNELLY

CHRISTINE COOK

Preventing High School Dropouts in Montgomery County Public Schools

Joshua Mausolf
University of Chicago
jmausolf@uchicago.edu

Bridgit Donnelly
University of Chicago
bdonnelly@uchicago.edu

Christine Cook
University of Chicago
ccook817@uchicago.edu

ABSTRACT

In this paper, we outline a machine learning approach to aiding Montgomery County Public Schools (MCPS) administrators in best targeting students most at risk for dropping out of high school.

In this paper, we approach the nationwide issue of high school dropouts, specific to the case study of Montgomery County Public Schools, a large school system in the Washington D.C. metro area. Utilizing data from this school district, we create a scalable ranking system that will allow MCPS to target their interventions to individuals most immediately at risk for dropping out of high school. Through our analysis, we show how techniques from machine learning provide a more effective tool for school administrators compared to their current methods. We suggest pathways for current policy improvement and additional avenues for future research.

We prepared the data by imputing missing values first on the individual level, then using the class means where individual imputation wasn't possible. We generated binary features to replace the categorical features and generated some new features of our own. We tested a variety of machine learning classifiers on the first cohort using K-Folds cross-validation and found that logistic regression and random forest gave us the best results. Using those two classifiers, we tested on the second cohort of data and evaluated our model using precision-recall curves and found that logistic regression was the best classifier in terms of both recall and precision.

Keywords

High School Dropouts, Educational attainment, Education Policy, Dropouts, Machine Learning

Index

1. INTRODUCTION AND MOTIVATION

According to the U.S. Census Bureau's Current Population Survey, high school dropout rates have steadily declined since the early 1990's, and as of 2013, only 7% of 18-to-24-year-olds drop out nationally (Fry 2014). Yet, even with these encouraging improvements, there are still 2,215,000 high-school dropouts in the 18-to-24-year-old population (Fry 2014). Beyond the stigmatized facade of being labeled a "high school dropout," this dropout population faces a myriad web of subsequent problems that impose long-term effects. To this point and to solve the problem of high school dropouts, we must first seek to understand (1) the effects of high school dropout, and subsequently seek to understand (2) the risk factors associated with high school dropouts.

1. 1 The Effects of High School Dropout.

Criminal Activity. For example, there is a known relationship between high school dropout and criminal activity. According to sociologist, Bruce Western (2006), high school dropouts are significantly more likely to find themselves behind bars compared to those with a high school degree, some college, or a bachelor's degree. Similarly, individuals in the bottom third income bracket remain those most often guilty of violent crime (Western 2006). Unfortunately, the economic returns of high school dropouts--on average--place them in this lower earnings echelon. Committing violent crime--or any serious crime for that matter--at a higher incidence than than the rest of the population further disadvantages high school dropouts on the job market. As Pager (2007) indicates, the mark of a criminal record irrevocably places felons at a disadvantage in pursuing employment. Moreover, this criminal mark disenfranchises felons from the political process by usurping their right to vote and thereby exercise their voice in American politics (Uggen and Manza 2002).

Economic Returns. Beyond the disadvantages levied in terms of criminal activity that increasingly befalls high school dropouts compared to others with higher educational attainment, economic returns for high school dropouts also suffer, even outside of the ostensible complications of a criminal record. The 2013 Current Population Survey indicates markedly lower incomes for high school dropouts compared to high school graduates. For 18-to-24-year-olds working full time, high school dropouts mean earnings were only $23,630 compared to $27,219 for high school graduates and $38,376 for those holding a bachelor's degree (Current Population Survey 2014). Beyond these ostensible gains for those with a high school degree, those with a bachelor's' degree earn even more, and even after adjusting for the cost of college, come out financially ahead of high school graduates (Hout 2012). Because high school dropouts are not even eligible for college without completing their GED, they unfortunately truncate their lifecourse possibilities at a very early age.

Wellbeing Benefits. Beyond these economic benefits, scholars have shown that high school graduates and those with some college or more are both markedly happier and healthier in their lives. For example, Hout (2012) illustrates that those lacking a high school diploma had a much lower incidence of reporting that they were "very happy" or had "excellent health" compared to their peers who graduated high school or pursued some other form of post-secondary education (394). In terms of family structure, households with higher education are more likely to be two-parent households. Whereas about 90% of college-educated households have two parents, only about 50% of households lacking a high-school diploma also have two parents (Hout 2012). For this reason, among others such as the complication with crime and economic returns, preventing high school dropouts remains an important national issue to resolve.

1.2 Risk Factors for High School Dropouts

Yet, regardless of the algorithm or technique used to prevent the nettlesome problem of high school dropouts, the practitioner must to some degree understand risk factors associated with dropping out in order to meaningfully intervene.

Academic and Behavioral Factors. With respect to what factors predict high school dropout, academic factors present perhaps one of the most salient factors. For example, studies have shown that a low GPA remains a powerful predictor of future high school dropout (Suh and Suh 2007; Coley 1995). Similarly, the incidence of a suspension or numerous absent days in school increases the likelihood that a student will drop out (Suh and Suh 2007).

Social and Neighborhood Factors. Beyond such overt problems such as subpar academic performance or behavioral delinquency, the social factors, particularly neighborhood effects can pose a lasting effect on students and their ability to complete school. For example, low household socioeconomic status can irreversibly diminish cognitive ability--even as early infancy (McLanahan et al. 2003). While our case study does not capture socioeconomic status at the time of infancy or early childhood, past socioeconomic status relatively correlates to current socioeconomic status, which in turn maps in aggregate to mean household income for a given census tract. Incorporating such data further thus relates a student’s academic abilities to the socioeconomic status of his neighborhood. Beyond individual effects of poor socioeconomic status, understanding the mean household income of the households surrounding a public school can be equally prognostic. For example, Diamond (2006) reveals a divide in the way school administrators encourage academic performance and correct for its deficiency both in terms of approach and resources. In almost every way, students at affluent schools outperform students at comparatively disadvantaged institutions (Diamond 2006). The “neighborhood effect” verily afflicts students within and can lead to not disadvantaged households and suffering schools but also higher proliferation of crime and compounding and concentrated cognitive disadvantage (Wilson 2012; Sharkey 2010).

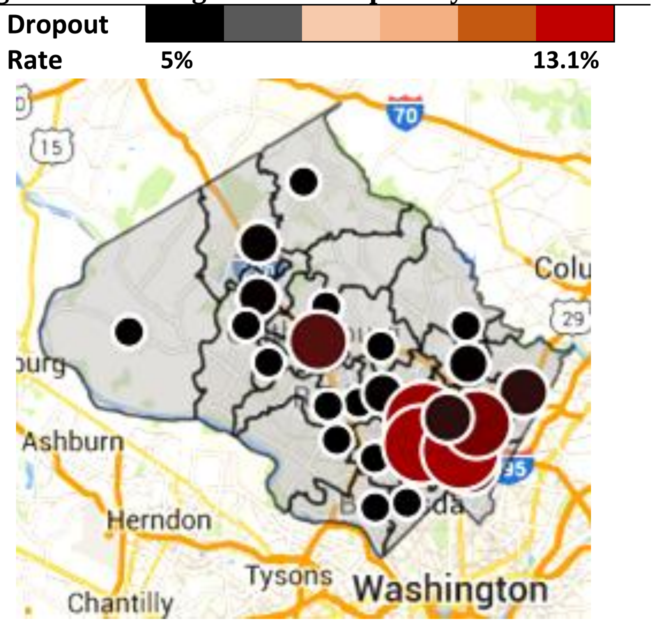
Thus, to condense and reiterate, the risk of dropping out drastically increases at an individual level for students not only academically or cognitively lagging but also those with behavioral issues. These problems do not originate in isolation--and in fact result in part following localized disadvantage at the neighborhood and social network level. Such social issues not only affect school resources, they affect family structure, adult mentorship, economic opportunity, and neighborhood crime--all of which can compound and either hurt or benefit individual academic performance--and high school graduation--depending on the directionality of comparative advantage and disadvantage.

1.3 Applying these Findings to Montgomery County Public Schools

To study this problem and develop an efficient model of predicting high school dropouts, we apply our techniques to the case of Montgomery County Public Schools (MCPS), a large school system in the Washington D.C. metro area. Although MCPS is one of the wealthiest school districts in the country, the cost of living is also among the highest--and this disparity yields the same pockets of advantage and disadvantage that affect high school dropouts. For example, while some high schools have dropout rates as low as 5%, others have dropout rates as high as 13.1%.

Our study attempts to incorporate not only the individual level risk factors that predict high school dropout, but also the school-level and neighborhood level (census tract-level) factors that help illuminate the problem of high school dropouts.

Figure 1 High School Dropout by School



Source: MCPS Open Data Portal

To provide the most detailed predictions of school dropouts, our study uses both individual level data from MCPS and school-level and neighborhood level data. With the data, we create a scalable ranking system that will allow MCPS to target their interventions to individuals most immediately at risk for dropping out of high school. Before turning to this analysis, we examine several problems with the current approach and illuminate the pathway for improvement.

2. CURRENTLY USED APPROACHES AND OUR SOLUTION

2.1 Problems with Current Approaches

Montgomery County Public Schools (MCPS), a large school system in the Washington D.C. metro area, is one school district interested in addressing this dropout issue. Currently, MCPS, which serves over 150,000 students, has an 88.3% graduation rate. MCPS developed their own dropout identification tool to target students at risk for dropping out, but they were interested in improving its accuracy by utilizing machine learning methods.

Currently, Montgomery County Schools uses a threshold-based approach that relies on warning indicators such as high absences or low GPA. The problem with this approach is that it is unscientific and expensive in terms of time and labor hours spent manually analyzing data on a case by case basis. Warning indicators are chosen anecdotally and the school doesn’t use historical data to predict future dropouts.

2.2 Our Solution

We aim to improve upon the current MCPS approach in three main ways: (1) leverage machine learning methods to take into account potentially predictive variables that are not currently incorporated in the threshold system; (2) build a flexible model that can expand or contract given the available resources in a given year; and (3) cut down on the amount of time required to prepare the model.

Our focus is on developing a model to predict dropout in 12th grade, the grade with the highest incidence of dropout. This model creates a sorted list of students in order of riskiness of dropping out, allowing teachers and administrators to focus their intervention efforts on those who need it the most.

Expected Policy Contribution

With the list of students organized by dropout risk, schools can offer targeted programming to keep these students in school. There is a lot of research on dropout prevention methods, the majority of which focus on engaging the parents, giving students specialized mentors, providing academic tutors and extra-help sessions, or offering evening classes to accommodate students who need to be working. Collectively, we hope that our analysis will not only increase the impact of current dropout interventions at MCPS, but that these tools can also be transferred to other schools wishing to assist those most at risk of dropping out.

3. DATA SOURCES

3.1 Existing Data

For this project, we acquired access to the Montgomery County Public Schools data for two cohorts of students, the 2012 cohort (Cohort-1), including years 2006-2012 and the 2013 cohort covering years 2007-2013 (Cohort-2). This dataset includes various features such as individual student IDs, gender, grade, school, PSAT scores, number of suspensions, absence rate, withdrawal codes, and whether they graduated on time, among other features. In total, there are 10,884 students in the last year of the 2012 cohort and 10,829 students in the last year of the 2013 cohort. Table 1 outlines the size of each cohort by grade and the number of dropouts in that year.

TABLE 1	**Cohort Size and Dropouts by Grade**			
	Cohort-1 Size (N)	**Cohort-1 Dropouts**	**Cohort-2 Size (N)**	**Cohort-2 Dropouts**
Grade 6	8681	11	8905	12
Grade 7	8995	4	9161	6
Grade 8	9300	2	9469	0
Grade 9	10299	30	10420	25
Grade 10	10708	96	10744	103
Grade 11	11001	242	10982	267
Grade 12	10884	315	10829	269

3.2 Supplementary Data.

In addition to the existing individual-level MCPS data, we also gathered supplementary (a) neighborhood-level data (at the census tract level) and (b) school-level data. Methodologically, we utilized the original MCPS data, which included the school name, to collect other facts about the school including its specific address and geocode.

With this critical information, we established the census tract for each school and used this data to collect additional data about the neighborhood containing the school, such as but not limited to the neighborhood mean income, the neighborhood mean income, the racial and demographic composition, the proportional highest level of educational attainment breakdown (by education level) for each neighborhood, and the proportion of youth population on food stamps, and the proportion of the youth population living in poverty. This census-tract level data originated from the 2012 American Community Survey.

For school-level data, MCPS has an open data portal that details school-level data such as the dropout rate per school, mean attendance per school, and the student-faculty ratio, among others. By supplementing the individual-level data with school and neighborhood data, we can best assess which features are most predictive of students dropping out of high school.

4. IMPUTATION AND FEATURE GENERATION

4.1 Feature Generation.

Defining High School Dropout. For this study, we based our definition of high school dropout utilizing the “withdrawal codes” or “wcode” from the MCPS dataset. This variable provided various numeric codes for a myriad of withdrawal reasons, some negligible or positive reasons such as transferring to school outside of MCPS, graduating MCPS, or graduating with honors. Alternatively, the withdrawal code entailed various negative forms of withdrawal such as dropout due to teen pregnancy, dropout due to loss of interest, illness, poor academic performance, academic expulsion, court-sanctioned incarceration, or simply that the student had unknown whereabouts, among other similar reasons.

As already intimated, a high school dropout was our dependent variable that we predicted. Yet, we were less concerned with the fact of dropping out at some point in high school than in pinpointing our prediction to which year they would drop out. Thus, we created dummy variables for dropping out at each grade level, and specifically predicted dropouts for grade 12, with potential future development around expanding to grades 10 and 11.

Additional Feature Generation. Regarding additional feature generation, we took several approaches. At one level, we conducted basic feature generation from existing data. We used strategies for example of creating binary and discrete dummy variables from categorical data. In this way, we maximized the algorithms ability to fully classify all available patterns. Beyond this base level feature generation, another level of feature generation results not from manipulating the existing data but by supplementing it with manifold neighborhood and school level vectors that shift with each student as they progress or flounder from school to school.

Extending these feature generation methods, we also manipulate the data to generate several new raw features, such as a feature identifying first the average GPA for a given year and using this information a secondary feature identifying the change in GPA from year to year. In this way, the model can detect improvement or decline in academic performance year over year.

4.2 Imputation Strategy.

To account for missing data in our dataset, we employed imputation to fill in these missing values. In particular, we utilized a "padding" technique of imputation, such that missing values were filled in by going "across rows" for a particular student. In this way, the missing values were conditional on the given student rather than the collective county-wide mean. Only in cases where this padding imputation was not possible due to sparseness of the data for a given observation did we rely on traditional mean imputation. Using this combination imputation strategy, we were able to gather the highest quality results without dropping observations from our dataset.

5. MODELS

In approaching this classification question, we tested six different basic classifiers, chiefly k-nearest neighbors, random forest, logistic regression, decision tree, decision tree with bagging, and decision tree with boosting. While the current method does not necessarily loop over parameters, we have customized our models as theoretically pertinent to our research question. For example, we limited our decision trees and random forest models to a maximum depth of five and our k-nearest neighbors to the three closest neighbors. In making the determination for which models to use, we conduct a robust training and testing schema using multiple methods of evaluation. We elaborate on both of these subjects below. We used class weights where possible to over-sample students who dropped out to prevent our model from predicting all 0's in years with very low dropout rates.

[THIS SPACE INTENTIONALLY LEFT BLANK]

6. TRAINING AND TESTING

Training and Testing Variable Setup. With respect to the actual variable implementation used in our analysis, regardless of the train/test splits, we are predicting on dropout in grade 12. For training the model, we only train on data for prior grade level years, namely data from grades 6-11.

Training and Testing Phases. For this analysis, we employed a robust training methodology. In the first phase of training, wherein we made choices regarding the final model selection, we exclusively trained and tested on the data from Cohort-1. During this phase of model training, we implemented five-fold cross validation. A key benefit of this technique was added robustness. In this way, we avoid "overfitting" the models. The imperative point of this strategy is designing a system that is sustainable for future and truly unknown test-data sets, as would be the case for a school system using this system. Thus, rather than the model only performing for the idiosyncrasies in Cohort-1, we have higher confidence that this final model will work not only for the Cohort-2 data, but also future cohorts in MCPS. Once the overall-top-performing model was selected from the first phase of training and testing, we retrained the model using all of the available data from Cohort-1 to test on Cohort-2.

7. EVALUATION METHODOLOGY

To evaluate these findings, we will focus on maximizing recall (the share of future dropouts we are able to successfully identify). In essence, this method emphasizes preventing any potential dropouts from slipping through the cracks. Secondarily, we focus on precision, especially relative to how the model performs in relation to various proportions of the population. Lastly, because only a small proportion of students in a given cohort will drop out in a given year (less than 10%) and because MCPS will likely not have funding to address every potential high school dropout, we looked to develop a model that would maximize both precision and recall specific to that 10% proportion of the population.

8. RESULTS

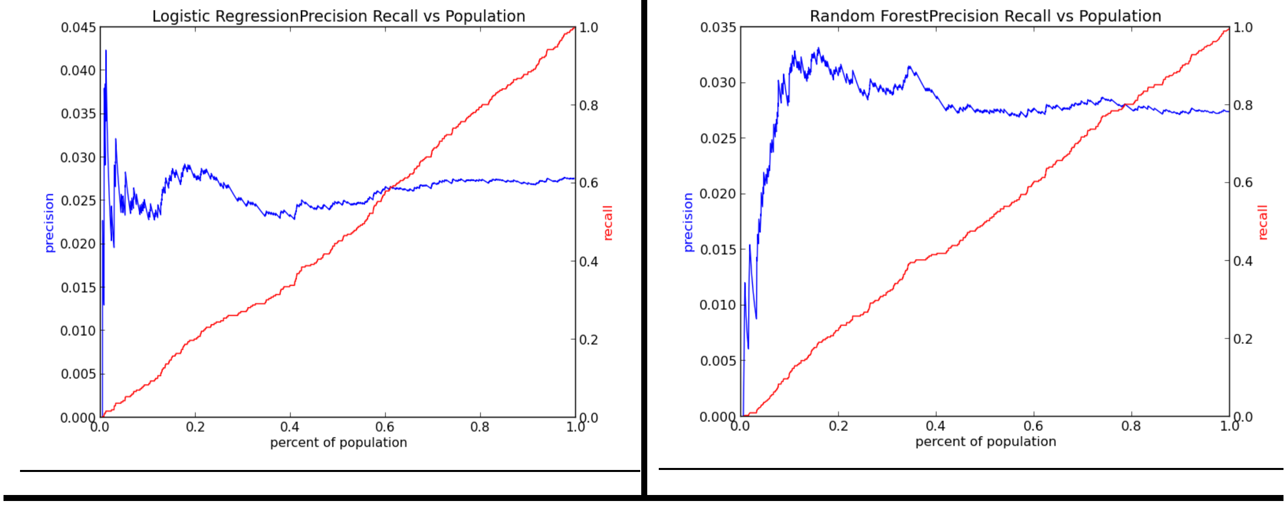
TABLE 2 **Evaluation Metrics for Chosen Models**

	Accuracy	Precision	Recall	F1 Score
Baseline	0.971			
Cohort 1 only: Logistic Regression	0.973	0.481	0.124	0.197
Cohort 1 only: Random Forest	0.972	0.299	0.184	0.228
Cohort 2 (trained on Cohort 1): Logistic Regression	0.978	0.474	0.152	0.231
Cohort 2 (trained on Cohort 1): Random Forest	0.967	0.231	0.225	0.228

After testing the six different basic classifiers (k-nearest neighbors, random forest, logistic regression, decision tree, decision tree with bagging, and decision tree with boosting). We found that both logistic regression and random forest maximized precision and recall in the 10% proportion of the population, as shown in Figures 2 and 3 (note that both axes are measured in decimal proportions rather than percentages). Additionally, both had accuracy scores above our baseline accuracy of 0.971 (Table 2)

FIGURE 2 **Precision and Recall Training and Testing on Cohort-1**

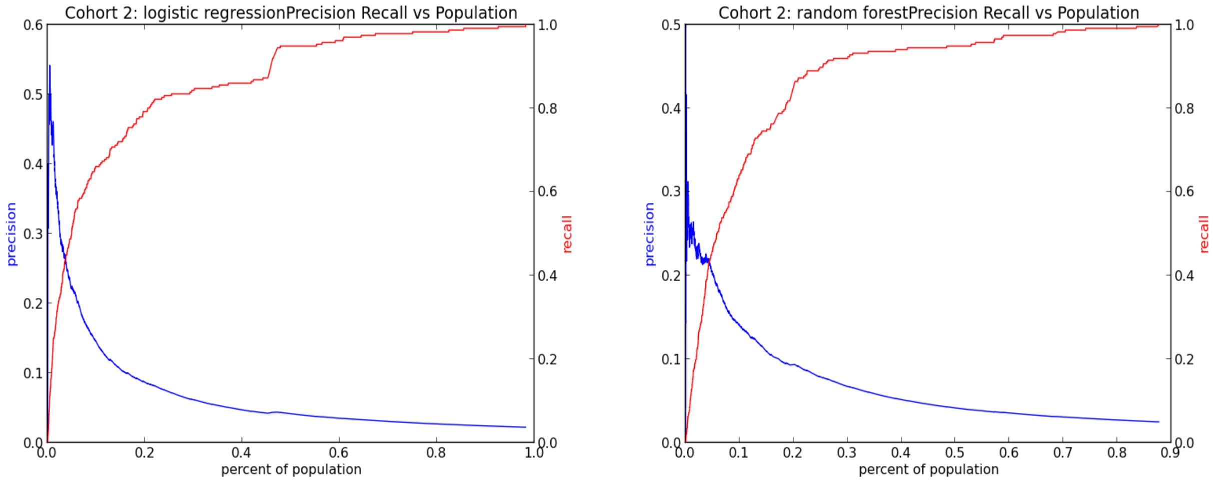
Logistic Regression **Random Forrests**



Since recall for all classifiers resembled a straight line, we looked to which classifiers had the highest precision, given 0-10% of the population was above a given threshold. When applied to cohort 2, we find the two models do reasonably well, as seen in Figures 4 and 5, and they continue to do well in terms of our other evaluation metrics.

FIGURE 3 **Precision and Recall Results for Testing on Cohort-2**

Logistic Regression **Random Forrests**



9. POLICY RECOMMENDATIONS

We would recommend using one of our two models to determine which students at risk of dropping out in 12th grade should be allocated resources in the coming year. In future years, we would encourage MCPS to be more diligent about filling missing data and collecting additional data. It would also be helpful for the schools to feed data back into the model at the end of the year to see if the model made accurate predictions.

10. DISCUSSION OF FUTURE WORK

10.1 Bolstering the Machine Learning Analysis.

Regarding future work on this project, there are several avenues for development. On one hand, future work will seek to bolster the rigor of machine learning. Practically speaking we will seek to generate additional features and further augment the machine learning pipeline. For example, we will strive to include additional machine learning algorithms to the pipeline and to loop over parameters for these models. We would also like to re-run our model to predict 10th and 11th grade dropouts in addition to 12th grade. Subsequently, we will analyze the results again in Cohort-1 to verify the two top-performing models that we recommend.

10.2 Improving Deliverables for School Administrators.

A natural progression of this work will be to design and develop a downloadable desktop application for school administrators. The overall design will be such that all the necessary machine learning modules and components are self-contained and “behind the scenes” for the consumer. The user-interface will consist of a basic, clean, and intuitive layout such that the consumer can drag and drop or otherwise use a drop-down menu to import existing cohort data that will comprise the training data. The consumer will be able to select several options such as the top number of students to target given the budget and cost per intervention that the user provides. The consumer will then click “generate results,” and after running, the program will deliver a polished spreadsheet illustrating the specific rank-ordered students to intervene with as well as other useful summary statistics about these students for the purposes of administrators presenting and discussing the results. Once the results for the predicted year are later known in the future, the user will be able to drag and drop additional spreadsheets so that the program can retrain with the additional information, and consequently, make improved predictions for students in the upcoming year.

11. ACKNOWLEDGEMENTS

First and foremost we would like to graciously thank Rayid Ghani and Gustav Larsson for their informative instruction, review, and guidance during our development of this project. We would also like to thank Data Science for Social Good of the Computation Institute at the University of Chicago as well as Montgomery County Public Schools for providing access to this data. Lastly, we would like to thank the Chicago Innovation Exchange for the generous use of their space and the much appreciated free coffee.

REFERENCES:

Coley, Richard J. (1995). *Dreams deferred: High school dropouts in the United States*. Princeton, NJ: Educational Testing Service, Policy Information Center.

Current Population Survey. 2014. "Table PINC-04. Educational Attainment--People 18 Years Old and Over, by Total Money Earnings in 2013, Work Experience in 2013, Age, Race, Hispanic Origin, and Sex." Washington, DC: United States Census Bureau.

Fry, Richard. 2014. “U.S. high school dropout rate reaches record low, driven by improvements among Hispanics, blacks.” *Pew Research Center*. Retrieved June 8, 2015 (http://www.pewresearch.org/fact-tank/2014/10/02/u-s-high-school-dropout-rate-reaches-record-low-driven-by-improvements-among-hispanics-blacks/).

Hout, Michael. 2012. “Social and Economic Returns to College Education in the United States.” *Annual Review of Sociology* 38(1):379–400.

McLanahan, Sara, Irwin Garfinkel, Nancy Reichman, Julien Teitler, Marcia Carlson, and Christina Norland Audigier. 2003. *The Fragile Families and Child Wellbeing Study: Baseline National Report*. The Bendheim-Thoman Center for Research on Child Wellbeing. Retrieved September 27, 2014 (http://www.fragilefamilies.princeton.edu/documents/nationalreport.pdf).

Pager, Devah. 2007. *Marked: Race, Crime, and Finding Work in an Era of Mass Incarceration.* Chicago: University of Chicago Press.

Sharkey, Patrick. 2010. “The Acute Effect of Local Homicides on Children’s Cognitive Performance.” *Proceedings of the National Academy of Sciences* 107(26):11733–38.

Suh, Suhyun and Jingyo Suh. 2007. “Risk Factors and Levels of Risk for High School Dropouts.” *Professional School Counseling* 10(3):297-306.

Uggen, Christopher and Jeff Manza. 2002. “Democratic Contraction? Political Consequences of Felon Disenfranchisement in the United States.” *American Sociological Review* 67(6):777–803.

Western, Bruce. 2006. *Punishment and Inequality in America*. New York: Russell Sage.

Wilson, William J. 2012. *The Truly Disadvantaged: The Inner City, The Underclass, and Public Policy*, Second Edition. Chicago: University of Chicago Press.