Bias in AP Computer Science Exam

Nanyu Zhang
Applied Data Science
University of Southern California
Los Angels, CA 90007
nanyuz@usc.edu

Junye Li
Applied Data Science
University of Southern California
Los Angels, CA 90007
junyeli@usc.edu

Abstract

AP Computer Science A is a course that contains the content of an introductory college computer science course. The goal of this project is to find out whether there exists gender and race bias in AP Computer Science A passing rate. The first approach is to calculate the difference between female and male passing rates which are predicted by the linear regression model. The second approach is to make some predictions with logistic regression model and calculate some fairness metrics such as statistical parity and equality of opportunity.

1 Introduction

Technology drives economic growth throughout the country. People are also starting to notice an imbalance in the number of men and women in the computer field. The data provided by National Center for Science and Engineering Statistics shows us there are $2,332,000$ male computer scientists and $796,000$ female computer scientists(7). There are three times as many male computer scientists as females. Also, we can see that there were $92,650$ graduate students with a computer science major in 2016. However, only 30 percent were women. These statistics show that computer science researchers are biased by gender. We believe that this difference exists because that gender difference is inherent in high school computer science education.

President Obama states that computational skill is one of the critical skill that the nation should empower students in order to thrive in the global economy in January 2016(10). Providing more basic computer science courses in high school teaching will not only address the basic educational needs but also the future labor market demand for computer technology personnel. Therefore, we need our next generation to end their K-12 education with more or less exposure to cs courses just like math. To find out if there are gender differences in students' exposure to cs courses in high school studies, we will focus on AP Computer Science A Exam. AP Computer Science A is a college-level introductory computer science course. AP Computer Science A exam expands students programming skills by solving problem using Java.

This project aims to find gender and race bias in AP Computer Science A Exam. We think average income, black people rate, white people rate, graduation rate of public high school, and research and development expenditure in different states are independent variables because we assume these variables will affect the passing rates. Firstly, we use the data from 2011 to 2018 to build a linear regression model. Then, we use this model to predict the average female and male passing rates among different states in 2019. We will calculate the difference between the predicted passing rate and the actual passing rate. By comparing these two differences, we can know if gender bias exists. Secondly, we use the data from 2011 to 2018 to build a logistic regression model. We will calculate some fairness metrics such as statistical parity and equality of opportunity to find out whether the gender and race bias exist or not.

[1]

[1] project data, code, Tableau plot: https://github.com/nanyuzhang/DSCI531-project

2 Literature Review

Bruno and Lewis(1) talk about whether high schools offer more computer science courses to every gender identity in California and whether every student regardless of their identity can access the same education resource. Firstly, they point out that the percentage of students be admitted into a high school that offers computer science courses has increased over the last 10 years in California. However, it still reveals a problem that Asian students enrol in more computer science courses than non-Asian, especially Hispanic, black and Native American students. In addition, the essay states that people with different identities do not have an equal chance to get the same education resources due to the lack of high-qualified computer science teachers. Also, computer science teachers are mostly white males. In this essay, the authors do not spend a lot of time analyzing the bias. In our project, we will try to use the definition of fairness to convince our hypothesis that there exists a bias in a secondary computer science course. Additionally, we will use the dataset from all 50 states.

Beyer and Sylvia predict females' computer science grades are more accurate than males'(2). They find previous computer science experience will not help people perform better. The research shows that factors that will increase females' grades will also improve males'. Also, it tells female students with higher grades always 'have positive stereotypes of computer science, have female computer science teachers in their high school and do not have gender discrimination in their department'(2). In our project, we will focus on finding the difference in AP computer science exams between females and males. However, this paper can help us know the cause of this difference.

In 'Gender differences in computer science students'(3), authors conduct research to find out the reason that leads to the lack of computer science major females. In their project, they sent a questionnaire to 56 students who are currently enrolled in one computer science course. They find that low confidence in computer science is an important factor that makes females not choose computer science. Stereotype threats and the perception of the academic environment are obstacles in the way of choosing a computer science major. However, I deem the sample size of this project to be too small. To know the reason, the research should be based on larger sample size. This article certainly gives us a good idea about what is the cause of such lack.

Cimpian (4) There is an obvious gender gap existing in some STEM majors especially for those math-intensive fields, such as engineering and computer science. The gender imbalance is related to the student's achievement level. These kinds of STEM majors attract more low-performing males and refuse low-performing women. There is a big gender difference between low-performing students and high performing students. The low performing male students have higher rate than the female low performing students. For our project, our dataset is about the AP CS exam. We will compare the difference between male students and female students by their AP scores and find whether the gender gap is related to students' performance in AP CS exams or not.

Geneseo (5) The gap between males and females from the computer science program is very obvious. The authors conducted a study to find the reason for this gender imbalance. The ratio of female students in the college may be higher than the male students, but the female students who take the introductory computer science course are much less than the male students. The female students who graduate from computer science major become lesser, which means there is a barrier for female students in cs major. One of the reasons for the gender gap in cs majors is retention, and another reason is the barrier for females to enter computer science programs. This paper used data from students in SUNY Geneseo, our dataset includes 50 states in the US. We are curious that we can get some findings with our AP CS exams datasets based on this paper's study.

Vandenberg (6) The authors conducted a study about computer science programs and relative activities for elementary students. The authors used two-way ANOVA and found that there is a statistical difference based on gender. The computer science attitudes for boys are much more than for girls. However, there is no significant difference based on race. We are inspired by this paper and also want to find the gender gaps and race differences for AP computer science exam. The sample size of this research is small (N=169) and also the ratio of the white student is a bit higher. Our dataset includes more samples, and our dataset includes 50 states in the US with different races.

3 Background

We use the dataset for AP Computer Science A exam. This dataset is collected by the College Board which is an organization that holds many standardized tests used by K–12 worldwide and these tests play an important role in the admission process(8). AP(Advanced Placement) exam is one of the exams held by the College Board in which high school students can learn some college-level courses(9). Therefore, we think this dataset can help us understand the bias in high school computer sciences education. There will be a total of 5 points on the AP exam. We count a score of more than three as passing the exam because many universities can redeem a score of more than three for a foundation course in that subject. In this dataset, we have the passing rate among all the states from 2011 to 2019.

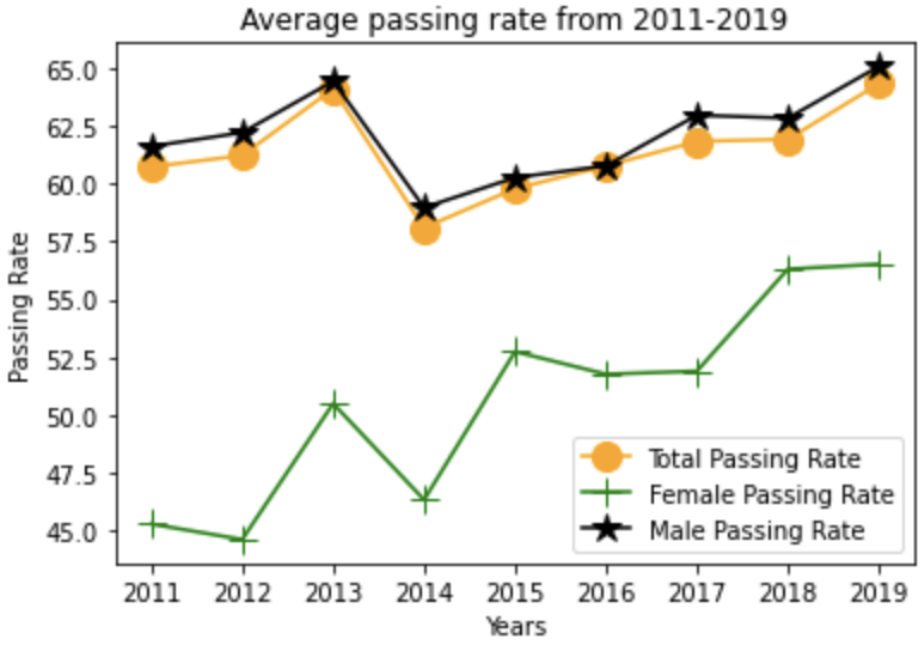

Figure 1

Figure 1 shows the passing rate of the AP Computer Science A Exam in the United States from 2011 to 2019. Surprisingly, the total passing rate has not been growing over time as expected. It shows that the total passing rate decreased in 2014. One possible reason is that the first question of the AP Computer Science A Exam in 2014 is very challenging. Therefore, I think many people do not know how to solve this question. Then, the total passing rate decreases by almost 7%. Also, we can see that the passing rate of female students increases from 45% to 55% between 2011 and 2019. The difference between female and male passing rates is shrinking. Additionally, the male passing rate is about the same as the total passing rate. The reason for this shape is almost 80% of people who take the exam are male.

Figure 2 shows the total passing rate of 51 states in 2011. Dark grey represents the state with a high passing rate, and light beige represents the state with a low passing rate(almost zero passing rate). We can see that the total passing rate in Mississippi, Montana, South Dakota, and Wyoming are zero in 2011 because only one or zero students attend AP Computer Science A Exam. According to our whole dataset, the passing rates of Montana and Wyoming in 2015 and passing rate of South Dakota in 2019 are zero because few people attend the exam. In 2011, there are 11 states whose female passing rates are zero. However, there are 4 states where total passing rates are zero. Since we find that the total number of students taking the exam and the passing rate are very low in Mississippi, Montana, South Dakota, North Dakota, and Wyoming from 2011 to 2019, then we will not predict the passing rate in the analysis.

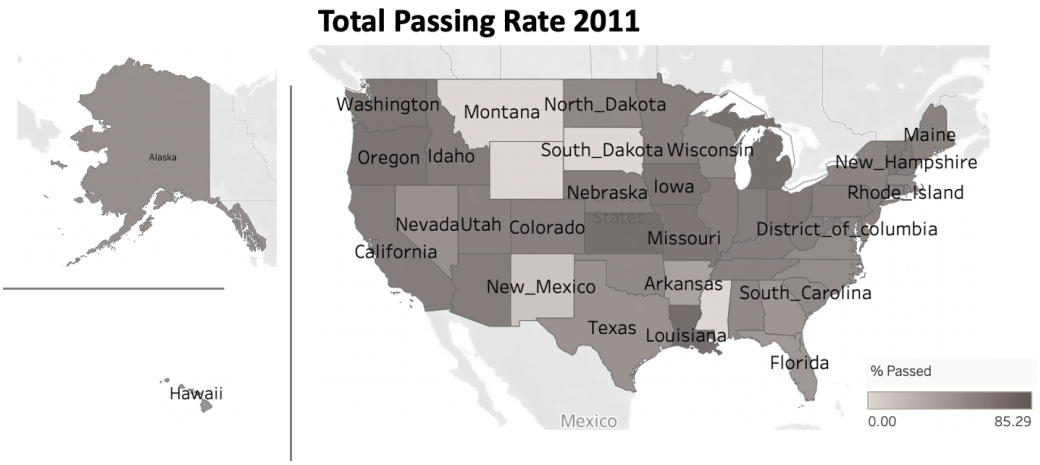

Figure 2

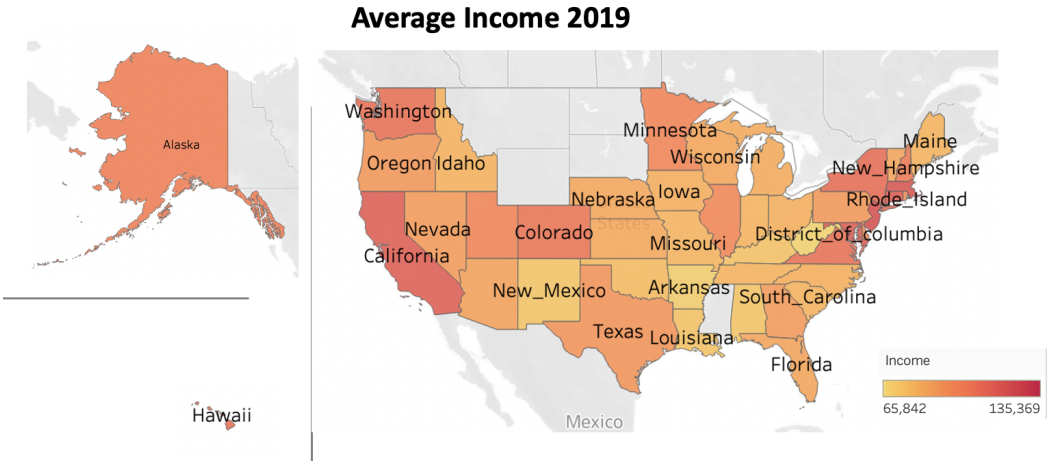

Figure 3

The educational attainment level of a state will highly impact the average income of a state. Nowadays, computing thinking becomes one of the critical skills that people should learn in this technology-based world. Therefore, we deem that a state with higher income should have higher educational attainment, especially in the computer science area. Then, we use the average income in 46 states(we delete five states that were mentioned before) between 2011 and 2019 from the United States Census Bureau.

Figure 3 shows the average income for each state in 2019. Dark red represents the high-income state, and yellow represents the low-income state. For example, the District of Columbia, California, and Washington are high-income states.

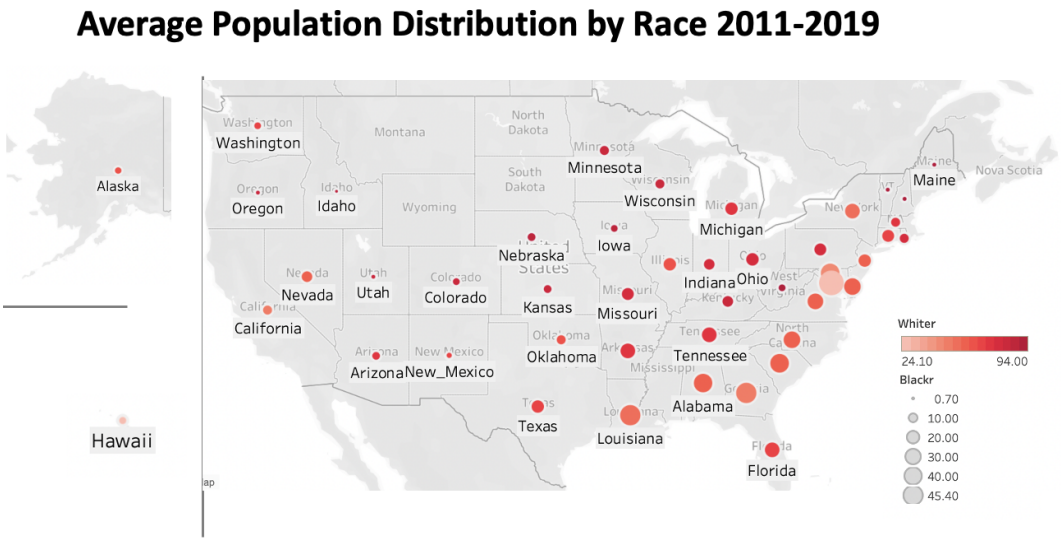

Figure 4

Another dataset we used is the distribution of the population by race for each state in the US from 2011 to 2019. This dataset is collected by the United States Census. We choose the proportion of the white people and the black people to represent the distribution of the population by race.

Figure 4 shows the average proportion of White and Black from 2011 to 2019 in 46 states(we delete five states that were mentioned before). Dark red represents the state with a high white population share, and light pink represents the state with a low white population share. The large circle represents the state with a high black population share, and the small circle represents the state with a low black population share. We can find that the distribution of the population by race is very different in each state.

Figure 5

The next dataset we used is the Federal obligations for science and engineering research and development to universities and colleges. This dataset is collected by National Center for Science and Engineering Statistics.

Figure 5 shows the federal obligations for science and engineering research and development to universities and colleges for 46 states(we delete five states that were mentioned before) in 2019. Dark blue represents the state with high federal RD obligations, and light green represents the state with

low federal RD obligations. We can find that some states such as California, Maryland, and New York are well above other states.

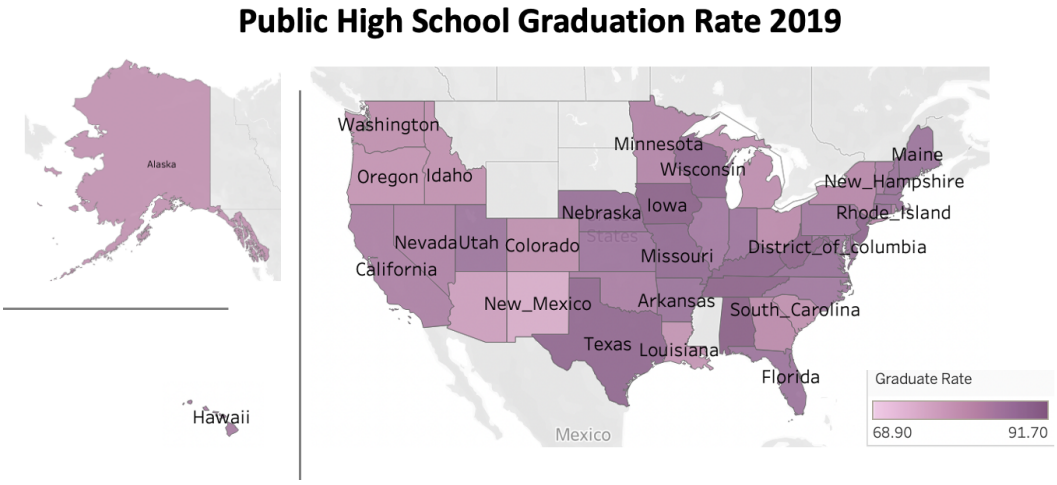

Figure 6

The last dataset we used is the public high school 4-year adjusted cohort graduation rate (ACGR). "The adjusted cohort graduation rate (ACGR) is the percentage of public high school freshmen who graduate with a regular diploma within 4 years of starting 9th grade"(11). We think this dataset can represent the high school education level for each state to some extent and can help us understand the bias in the high school computer science education.

Figure 6 shows the Public high school 4-year adjusted cohort graduation rate for 46 states(we delete five states that were mentioned before) in 2019. Dark purple represents the state with a high graduation rate, and light purple represents the state with a low graduation rate. The graduate rate for most states reaches 80%.

4 Method

As we mentioned before, we decided to delete the five states(Mississippi, Montana, South Dakota, North Dakota, Wyoming) data because very few students attend AP CS A Exam in these states and the total passing rates of these five states are zero. We use data from 2011 to 2018 as the training set to predict the female passing rate and male passing rate in 2019. We choose linear regression as our regression model to predict the passing rate. Our independent variables are average income, the black population share, the white population share, the graduation rate of public high school, and federal obligations for science and engineering research and development expenditure.

	Income	whiter	blackr	graduate_rate	RD_expenditure	PassRate	FemPassRate	MalePassRate
Income	1.000000	-0.428750	0.152657	0.081933	0.358425	0.203096	0.162977	0.213139
whiter	-0.428750	1.000000	-0.573191	0.344759	-0.258642	0.319923	0.223098	0.296799
blackr	0.152657	-0.573191	1.000000	-0.225774	0.148540	-0.247797	-0.194452	-0.221587
graduate_rate	0.081933	0.344759	-0.225774	1.000000	0.061969	0.123650	0.293228	0.106040
RD_expenditure	0.358425	-0.258642	0.148540	0.061969	1.000000	0.160761	0.210318	0.171999
PassRate	0.203096	0.319923	-0.247797	0.123650	0.160761	1.000000	0.455518	0.978858
FemPassRate	0.162977	0.223098	-0.194452	0.293228	0.210318	0.455518	1.000000	0.365046
MalePassRate	0.213139	0.296799	-0.221587	0.106040	0.171999	0.978858	0.365046	1.000000

Figure 7

We calculate the correlation between each feature. According to Figure 7, we find that the correlations between each independent variable are very low, which avoids collinearity. We use our model to predict the female passing rate and male passing rate respectively.

Furthermore, we want to find out whether the bias exists when predicting the passing rate by race and gender. We will make a prediction using the logistic model first, and then the fairness metrics such as statistical parity and equality of opportunity. Because the dependent variables, with the exception of state, are all numeric variables, we will create 3 variables first. Given that the passing rate is numeric, we will convert it to a binary variable. If the passing rate is higher than the average passing rate, we will label it as 1. Otherwise, the value will be 0. Also, we need two protected groups to present the gender and locations respectively. To begin, we find that the male passing rate is highly correlated to the total passing rate. We calculate the male participation rate. Surprisingly, the average male participation rate is approximately 83%. Then, we generate a new variable by calculating if the percentage of male students who took the exam surpassed the average male participation rate. If yes, the value will be 1. Otherwise, it is a 0. Similarly, we create a variable by determining whether the percentage of white people in a state exceeded the average white rate. After changing these three numeric variables to categorical variables, we use the data from 2011 to 2018 to be the training set and predict the outcome of 2019. Then, we'll utilize statistical parity and equality of opportunity to see if there is a bias.

5 Analysis

We calculate the mean square error(MSE) to compare the female passing rate and male passing rate. According to Table1, we can find that the MSE for male students is much lower than the MSE for female students. The prediction for male students is much better than the prediction for female students, which means that gender bias exists.

Table 1: Passing Rate MSE

Male passing rate MSE	**Female passing rate MSE**	**Total passing rate MSE**
58.83	146.48	63.90

Table 2: Fairness with Gender

Accuracy	**Statistical Parity**	**Equality of Opportunity**
0.72	0.195	0.083

Table 2 shows that the accuracy of our model is 0.72. This implies that our model is good. Since the statistical parity is 0.195 which is not equal to zero, then there is a bias when predicting states with varying male participation rates. It shows that the probability of a state with higher male participation rate will have a higher chance to get a higher passing rate than that of a state with lower male participation rate. The equality of opportunity tells that the classifier is more likely to assign a high passing rate to a state with higher male participation rate and its actual passing rate is high, compared to a state with lower male participation rate. As a result, we may conclude that the bias exists when we predict the passing rate by different gender.

Table 3: Fairness with Race

Accuracy	**Statistical Parity**	**Equality of Opportunity**
0.72	-0.072	-0.105

Table 3 shows that the accuracy of our model is 0.72 which is the same as the previous model. This implies that our model is good.Although the statistical parity is -0.072 which is approaching zero, it is not equal to zero. Then, there is a bias when predicting states with different white rates. It shows that the probability of a state with higher white rate will have a lower chance to get a higher passing rate than that of a state with lower white rate. The equality of opportunity tells that the classifier is less likely to assign a high passing rate to a state with higher male participation rate and its actual

passing rate is high, compared to a state with lower male participation rate. Despite the fact that these statistics contradict our previous predictions, we still can conclude that a bias occurs when we anticipate the passing rate by race.

6 Conclusion

In this project, we made a prediction by linear regression model and calculated the difference between the predicted passing rate and the actual passing rate for males and females. We found that models can predict male passing rate more accurately. Then, we made predictions by logistic regression model and calculated some fairness metrics to find the gender and race bias. We found that if a state has a higher percentage of men participating in the AP CS Exam, a better prediction is obtained. If a state has a higher white rate, the prediction will be worse.

References

[1] Bruno P, Lewis CM. Equity in high school computer science: Beyond access. Policy futures in education. Published online 2021:147821032110630-. doi:10.1177/14782103211063002

[2] Beyer, Sylvia. (2008). Predictors of Female and Male Computer Science Students' Grades. Journal of Women and Minorities in Science and Engineering. 14. 377-409. 10.1615/JWomenMinorScienEng.v14.i4.30.

[3] Beyer, Sylvia Rynes, Kristina Perrault, Julie Hay, Kelly Haller, Susan. (2003). Gender differences in computer science students. ACM Sigcse Bulletin. 35. 49-53. 10.1145/792548.611930.

[4] Cimpian, Joseph R., et al. "Understanding Persistent Gender Gaps in STEM." Science, 19 June 2020, www.science.org/doi/full/10.1126/science.aba7377.

[5] Geneseo, Greg Scragg SUNY, et al. "A Study of Barriers to Women in Undergraduate Computer Science." ACM Conferences, 1 Mar. 1998, dl.acm.org/doi/10.1145/273133.273167.

[6] Vandenberg, Jessica, et al. "Interaction Effects of Race and Gender in Elementary CS Attitudes: A Validation and Cross-Sectional Study." International Journal of Child-Computer Interaction, Elsevier, 6 Apr. 2021, www.sciencedirect.com/science/article/pii/S2212868921000283.

[7] (2019). Retrieved February 1, 2022, from https://ncses.nsf.gov/pubs/nsf21321/data-tables.

[8] Wikimedia Foundation. (2022, February 1). College Board. Wikipedia. Retrieved February 2, 2022, from https://en.wikipedia.org/wiki/College_Board

[9] What is Ap? What Is AP? – AP Students | College Board. (n.d.). Retrieved February 2, 2022, from https://apstudents.collegeboard.org/what-is-ap

[10] National Archives and Records Administration. (n.d.). Computer science for all. National Archives and Records Administration. Retrieved March 25, 2022, from https://obamawhitehouse.archives.gov/blog/2016/01/30/computer-science-all

[11] Coe - Public High School graduation rates. (n.d.). Retrieved May 2, 2022, from https://nces.ed.gov/programs/coe/indicator/coi