The Impact of High School Region Socioeconomic Status on Computer Science Student Performance

Jennifer Alexandra Thompson
Virginia Tech
Blacksburg, Virginia, USA
alexthompson06@vt.edu

Margaret Ellis
Virginia Tech
Blacksburg, Virginia, USA
maellis1@vt.edu

Sara Hooshangi
Virginia Tech
Blacksburg, Virginia, USA
shoosh@vt.edu

***Abstract*—Research in computing education has been steered towards understanding early indicators of what leads students to succeed in introductory programming courses (CS1). A major finding of these research efforts has been the impact that high school courses and prior programming experience have in predicting success in a post-secondary CS1. However, the socioeconomic status surrounding CS1 students has not been well explored as an indicator of performance. Specifically, a student's high school socioeconomic status (SES) has not been well investigated in this area, despite the intuition that more socioeconomically advantaged high schools will better prepare students for college computing courses. In this research, we propose a method to examine a student's prior high school regional socioeconomic status and determine whether this SES has a correlation to their post-secondary CS1 performance. This paper investigates the socioeconomic status of the neighborhood, census tract, and county the high school resides. To understand the socioeconomic statuses of these regions, we utilize multiple socioeconomic indices such as the Area Deprivation Index and the Social Deprivation Index. Some of the factors that create a deprivation index are the housing values of the region, poverty rate, adult educational completion, and household resources. After proposing a method to examine if there are any correlations between a student's attended high school regional SES and the student's performance in CS1, we perform a case study using seven years of CS1 student records from our institution. From the 4863 student records we use in this study, our initial findings indicate that students from more advantaged high school regions tend to pass CS1 more frequently across all surrounding region sizes we examined. Since our findings indicate that high school regional socioeconomic status may be a factor in a student's performance, we argue that future computing education researchers should consider a student's SES as a demographic factor of course performance in order to advocate for interventions that mitigate this disparity gap.**

***Index Terms*—computer science education, CS1, high school preparation, socioeconomic status, quantitative analysis**

I. Introduction

Learning and applying the fundamental principles of computer programming has been shown to be a difficult task for new learners [1]. As a result, many research studies have been conducted on factors that influence student performance in introductory computer science (CS) courses. For students who take introductory programming courses at a post-secondary institution, there is a strong tendency to view students' high school courses as potential indicators of their success in CS courses. Many studies have shown that both prior programming experience and math courses completed in high school can be predictors of performance in CS [2], [3]. Additionally, some studies have indicated that English achievement and communication skills can also contribute to doing well in CS [4]. In our previous research, we have also observed Advanced Placement (AP) course credits as a possible indicator of success in post-secondary computer science classes [5].

While prior computing and math courses have been shown to be correlated with introductory CS1 performance, another important factor that could have a potential impact on student's success is the regional socioeconomic status of their attended high school. Well-funded public high schools are typically located in neighborhoods with a high median household income. Better-funded schools have more resources to introduce students to the computing field and can provide structural support to elevate students' interest [6], [7].

Presently, the United States does not have a standard socioeconomic status measure used for education purposes, and there is not a single standard index for measuring the quality of public or private high schools in the United States. This study will seek to fill this gap by presenting a methodology that can be used to numerically measure the socioeconomic status (SES) of the area surrounding a high school. Using this measure researchers can determine whether such socioeconomic standing has any impact on student performance in introductory CS courses. We will also argue the importance of considering socioeconomic factors as an indicator of success in education, which will be supported by the findings of this work.

This work is a continuation of a poster that investigates a correlation between student performance and one of the three socioeconomic indices we use in this paper [8]. This paper will continue this research by integrating the use of two more socioeconomic measures and taking a closer look at student performance by closer examining SES status between different performance groups.

In the next section of this paper, we discuss prior research on the impact of high school courses on CS1 performance, socioeconomic factors that relate to educational quality, and the current use of various indices to measure socioeconomic status. In Section 3, we describe some of the terminology used throughout this paper, like SES regions, identifiers of these regions, and the three socioeconomic indices used in this study. Section 4 lists the research questions we plan to answer in this study which focus on the correlation between of a student's high school SES and their performance in CS1. In Section 5, we propose a method to determine a high school's surrounding regional socioeconomic status using a combination of three indices and outline a metric for student performance. To test this method, in Section 6, we describe a case study using seven years of CS1 student data at our institution. We showcase our findings and address our research questions in Section 7. In Section 8, we will discuss the correlation between student performance and corresponding high school region socioeconomic status based on our case study. Lastly, we will discuss the current limitations of this research in Section 9 and its potential future use in educational contexts in Section 10.

II. Background

This section will highlight prior research on the effects of high school quality and resources in predicting success in post-secondary introductory computing courses, and examine the correlation between education quality and socioeconomic status.

A. High School and CS1

Factors that influence retention in introductory CS classes, in particular CS1, have been widely researched by educators to help understand where students struggle. Finding indicators of CS1 performance will help post-secondary educators understand educational gaps and assist the development of strategies to further strengthen the learning outcomes in courses.

A prevalent indicator of CS1 performance is the quality of classes taken in high school and students' grades in high school. Within the high school experience, there are many factors, such as prior programming experience, mathematics courses, and AP courses that could be examined. Horton and Craig saw that in their CS1 classes, students scored 24% higher on exams in CS1 if they had taken a computer programming course in high school as opposed to having no prior programming experience [4]. Additionally, the number of math courses taken in high school may affect a student's college-level computer science performance. Wilson and Shrock presented that the number of high school math courses students take could be a strong predictor of student success in CS1 [3]. The number of AP courses taken by a student in high school may also have a large impact on the student's performance in CS1. Alavardo et al. showed that students who had any AP credits may have higher grades across many college-level computer science courses [9].

Some research has been done on establishing the link between CS1 performance and a student's socioeconomic status. Parker et al. found inconclusive results to tie computer science performance to SES, despite literature in other disciplines correlating academic performance and socioeconomic status [10]. Instead, Parker et al. found that spatial skills could be tied to a student's computer science academic performance regardless of the student's socioeconomic status. This study seeks to re-investigate the hypothesis made on socioeconomic status and computer science performance.

B. Educational Quality and Socioeconomic Status (SES)

The socioeconomic status (SES) of a student plays a significant role in determining the quality of education they have access to and the post-secondary opportunities they can consider [11]. A student's socioeconomic background can determine the tutoring services they can receive, whether the student can afford to attend post-secondary education, and what level of educational attainment they receive as a result. The National Center for Education Statistics reports in a case study that students with a higher SES are four times more likely to finish a Bachelor's degree compared to students who come from a lower SES. This study constructed the individual students' SES from their parents' occupation, the highest level of educational attainment achieved, and household income [12]. In a recent study, students in a CS2 course who received financial aid only had a 69.8% success rate, compared to the 84.8% success rate for students who received no financial aid [13]. There is also a disparity in educational success among school districts with varying economic funding. One study shows that students who take an AP course are two times more likely to pass the course if they come from a high school in the top 25% of affluent school districts as opposed to the bottom 25% of poorer districts [14]. Students' parents' occupations, the highest level of education, and income aid information is not easily accessible, but students' high school name and address may be part of many student records.

Currently, there is not a single standard index for measuring the quality of public or private high schools in the United States, nor is there a single source of publicly available data on the funding levels of individual high schools and the funding per pupil. However, some private investigations have been used to determine the quality of public United States high schools. U.S. News and Research Triangle Institute (RTI) International research group have created a methodology to rank public high schools in the United States and to identify top-performing public high schools on a yearly basis [15]. The latest iteration of the ranking presented on the U.S. News website for the year 2022 has ranked nearly 18,000 schools on six factors including AP/IB exam scores, standardized testing, underserved student performance, and graduation rates. While this group did publicly release a paper on their methodology [15], they did not publicly release the raw score data for each high school. Instead, the website only lists the schools in the order of their ranking to determine the top schools in the country [16].

Given the limitation of the above ranking model and the lack of other publicly available resources to measure high school quality and its impact on student success in computing, in this work, we propose a new methodology to map a high school surrounding neighborhood to well-established socioeconomic indices to determine the local SES. The reason for choosing this measure is that some studies have shown that school districts with higher poverty levels often receive a lower amount of public school funding [17]. Thus, we hypothesize that the region surrounding the high school will likely give us an indication of the level of funding available to the school and subsequently the level of resources available to the students at that particular school.

C. Socioeconomic Deprivation Measures

In this work, we will use several regional socioeconomic deprivation indices to determine the SES of a student's high school region. A socioeconomic deprivation measure is a measure based on education, wealth, social, and economic factors of a given regional section of a country. The index is typically a national percentile between 1 and 100, where having a deprivation index of 1 means being in the top 1% of socioeconomically advantaged areas in the country.

Socioeconomic deprivation measures are currently primarily used in the medical field. These measures have been used to determine the life expectancy of patients, recovery rate, use of medical care, and populations with mental health illnesses [18]–[20]. Presently, the United States does not have a standard measure for socioeconomic deprivation, unlike countries such as the United Kingdom and New Zealand [21]. However, research institutions and organizations in the United States have created their own socioeconomic deprivation indices [22]–[24]. Pampalon et al. note the rationale for creating deprivation indices is to "illustrate social inequalities in health and the use of health services" [18]. While Pampalon et al. argue that deprivation indices can showcase healthcare inequalities, we are interested in investigating if these measures can present socioeconomic inequalities in education as well.

III. Terminology Used

This section will introduce socioeconomic deprivation measures and explore some of the terminology used throughout this paper, including the three deprivation indices we will use in this work.

A. Counties, Census Tracts, and Census Block Groups

In this study, we use socioeconomic deprivation measures, based on different regions surrounding the high school they cover. Since a socioeconomic deprivation measure evaluates a certain federally-determined region size, we chose to investigate multiple region sizes surrounding the high school. Ideally, we would have liked to look at school district-determined regions and school boundary areas. However, this data would be difficult to find based on how individualized and shifting these regions are for each school district. Instead, we focus on looking at federal regional measures including census block groups, census tracts, and counties.

Figure 1 illustrates how these federal regions are defined and are related to each other. Each of the 50 states in the United States is divided into smaller divisions called counties. Counties are large administrative districts that vary in size. Each county is divided into census tracts (CT) and each census tract is divided into several census block groups. Census tracts typically contain approximately 4,000 residents, while census block groups usually contain 250 to 550 housing units [25]. Census tracts are somewhere between twice and four times as large as census block groups, but this varies widely. Additionally, census block groups are often defined as neighborhoods in the context of socioeconomic measures [23], [24].

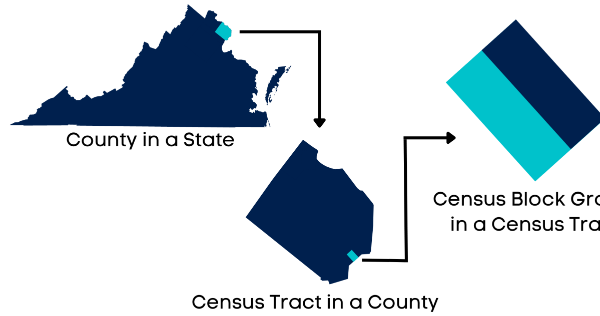

Fig. 1. A visualization of the relative sizes between states, counties, census tracts, and census block groups.

B. Census Block Group Code

Since each of these regions is contained within one another and is used for similar administrative purposes in the United States, there are common identifiers used to recognize a county, census tract, and census block group for a specific location. These geographical ID codes (GEOIDs) are defined by the United States Census Bureau and a GEOID that describes the county, census tract, and census block group of location is a Census Block Group Code [26]. The Census Block Group Code is a 12-digit numerical code and is broken down into several segments as shown in Figure 2.

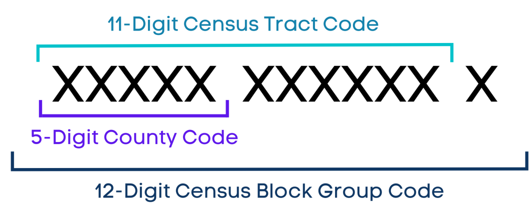

Fig. 2. A breakdown of how the Census Block Group Code is composed. The first five digits of the code uniquely identify the county, the first eleven digits represent the census tract and the full 12-digit code identifies the census block group.

The first two digits of the number uniquely define the state and together with the following three numbers, create a unique 5-digit identifier of a county in the United States. The first 11 digits of the Census Block Group Code identify the exact census tract, and the full 12-digit code identifies the census block group for a location [26]. With the full 12-digit code, we have the county, census tract, and census block group identifiers needed to use the ADI and the two SDI deprivation indices. Therefore, as part of our method, we will identify this code for every student's attended high school.

C. Area Deprivation Index

The first socioeconomic index measure we used in this study is the Area Deprivation Index (ADI). ADI measures socioeconomic status on a United States census block group and defines SES over seventeen different socioeconomic factors, such as level of education completion, employment, income, poverty rate, and household characteristics [23], [24]. ADI is divided into two index measures: a state decile which compares the block group to the rest of the state and a national percentile which compares the block group to the rest of the United States. Since students entering a post-secondary institution may come from different states across the nation, we decided to solely focus on the national ADI measure in this study.

D. Social Deprivation Index - Census Tract & Counties

The next two socioeconomic indices we use in this study are created from the Social Deprivation Index (SDI) and cover two differently sized regions surrounding the high school: census tracts and counties [22]. The SES defined by SDI is based on seven socioeconomic factors including years of educational completion, poverty rate, and household characteristics. Since we need a Census Block Group Code for each school to determine the census block group for our ADI measure, we decided to look at the SDI measures for census tracts and counties, as census tract and county numbers are contained within the Census Block Group Code.

Additionally, we want to look at census tracts, which are approximately twice the size of census block groups, to gain a more general understanding of the area surrounding the high school. Comparing neighborhoods represented by ADIs to larger regions will help us compare if findings on student performance are consistent when compared with socioeconomic indices that cover larger regions.

Lastly, we examine counties to best estimate school districts and larger area trends. Additionally, funding for schools may be determined on a county level which may equate to a trend of performance in students. However, given the larger overarching area a county regards, we do hypothesize these results may be less accurate overall than the census tract and census block group socioeconomic measures.

IV. Research Questions

This research seeks to understand what (if any) correlation exists between a CS1 student's prior high school area SES and their performance in CS1. To do this, our research questions compare students' CS1 passing rates and performance to their corresponding high school region SES as stated below:

- **RQ1**: Are students more likely to pass CS1 if they come from a more advantaged high school region socioeconomic status?
- **RQ2**: Are students more likely to receive a higher grade in CS1 if they come from a more advantaged high school region socioeconomic status?

V. Method

In order to determine if there is a correlation between a high school region SES and a student's performance in a computer science course, we propose a method to match a student's high school to three socioeconomic index measures of three differently sized surrounding areas. With this information, we then look at the students' final grades in the computer science course and determine if there are any trends between the socioeconomic advantage of a surrounding high school region and student performance. While in our case study, we look at CS1, we propose that this method could be used with any post-secondary educational course.

In this section, we describe how to match students' high school information to socioeconomic indices based on student record information. Then, we look at why we identified the passing rate of CS1 as a metric for performance. We will also touch on how to correlate the performance measure to an SES. After outlining our method in this section, we explore utilizing this method in a case study in the following section.

A. Matching Student High School Information to Socioeconomic Indices

After determining the indices we will be using to measure socioeconomic deprivation in the region of a student's high school, we need to use the information given in student records to determine correlating socioeconomic measures. This process of matching the student's high school information to socioeconomic indices is illustrated in Figure 3. Our steps here will be the following:

1) Match each high school name and zip code to an address.
2) Match each address to a Census Block Group Code.
3) Match each Census Block Group Code to a:
    - Census Block Group ADI
    - Census Tract SDI
    - County SDI

The student records we had accessed only had the high school name and ZIP code. From here, we needed to obtain the addresses of each of the high schools to geographically pinpoint its location. To determine the high school addresses, we used a master list of all high schools in the United States with College Entrance Examination Board (CEEB) codes. These codes are assigned to nearly every secondary school in the United States and are designed to report schools for SAT entrance exams [27]. The CEEB database we used listed information such as the high school's address, school type (public, private, etc.), and CEEB code [28].

From the addresses of high schools students attended, we then took the Census Block Group Code from each of the high schools to determine the block group, census tract, and county information. To do this for a large data set, we used the United States Census Bureau Geocoder API [29] to find the Census Block Group Codes of each high school address obtained from the CEEB search. Any information that was not recovered from the CEEB master list or the Geocoder Tool was manually determined through search engines.

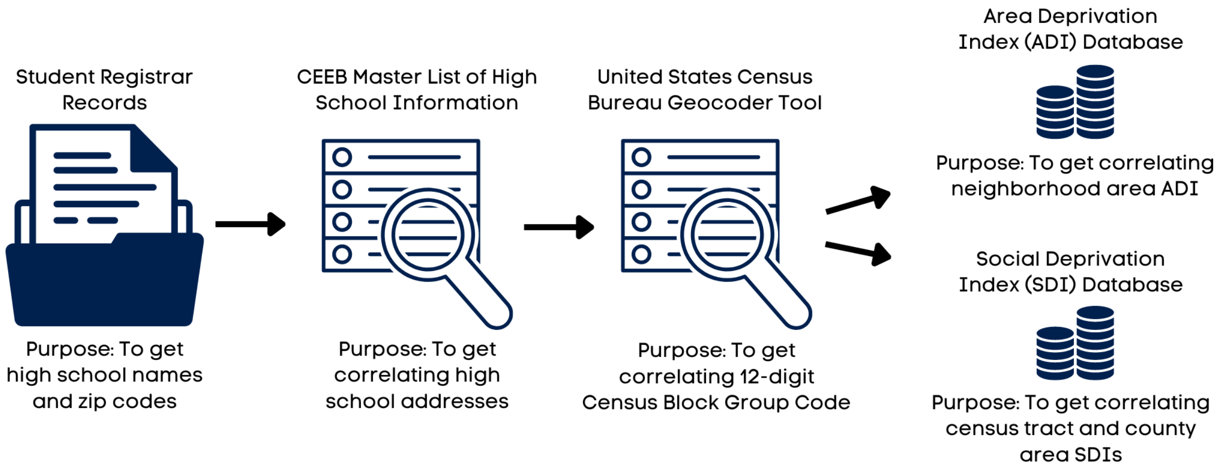

Fig. 3. A visual on our method of retrieving student high school information from registrar records to determine the three deprivation index values of the high school region.

Finally, after obtaining a 12-digit Census Block Group Code for each high school, we could determine the corresponding census block group ADI, census tract SDI, and county SDI. We utilized the publicly available data associated with each of these measures [22]–[24] with the corresponding portion of the Census Block Group Code needed to uniquely determine the county, census tract, and census block group that the high school is in. With this, each high school that students had attended was matched to an address and a Census Block Group Code. From there, we were able to match each high school to three socioeconomic measures concerning the region in which the high school resided.

B. CS1 Passing Rate & Performance Groups

With the corresponding socioeconomic data found for each student's high school region, we look for correlations between the socioeconomic information of a student's high school and the student's performance. For measuring performance in CS1, we use the overall final CS1 grade as the performance metric to determine the passing rate and performance groups of the students, according to our two research questions.

Passing rates are an important indicator of success, as they determine if students will be able to move forward with taking a CS2 course in post-secondary institutions. The passing rate for CS2 eligibility has also been used in other studies that look for predictors of CS1 performance [30]. In this study, we define the passing CS1 group as the group of students who successfully pass CS1 with a sufficient grade for enrollment in CS2. Alternatively, the non-passing group is those students who receive an insufficient grade for enrollment in CS2, or those who withdraw from the course. The use of identifying this passing rate will help us answer RQ1.

With the CS1 passing information gathered, we divided the population into sub-populations (referred to as splits in the rest of this paper) for each of the SES measures to determine if certain groups pass more often than others. This method was inspired by several medical studies that utilize socioeconomic deprivation indices, as they tend to split their population into different groups based on SES percentile to determine performance differences between multiple different socioeconomic statuses [19], [20]. We examine each of the three deprivation indices (ADI, SDI census track and SDI county) separately in this study, to determine if region makes a difference in the trends we see between different splits.

To address RQ2, we separate students into performance groups to get a more fine-grained look at the students from different performance levels. We identified a total of five performance groups in this study. We divided the students into excellent performance group (A to B+), average performance group (B to C), and poor performance group (C- to D-), as well as categories for students who failed and withdrew from the course. At our institution, in order to enroll in CS2, a student must have at least a C in CS1. This means that students from the excellent and average performance groups will be able to enroll in CS2 and students from the poor performance group will not. However, students from the poor performance group will still receive credit for the course, unlike those who fail or withdraw.

VI. Data

We performed the methodology described above with seven years of student records from our institution, Virginia Tech. This study is a continuation of a previous work, in which we looked at factors that influence student success in introductory CS classes [5], [31]. We examined the trends between the students' passing rates in CS1 and their reported high school region SES using ADI and SDI measures.

We acquired seven years of student records after IRB approval from the University Registrar's Office at Virginia Tech, a large research-focused university in the United States. This data contains records of CS1 student grades from Fall 2015 to Spring 2022. Each record includes a student's course information, final grades, demographic information, previous high school information, and majors. We filtered the data set to only focus on students who attended a high school in the United States, as the deprivation measures we use are calculated on a US national level. We found that 4863 students in this data set attended a high school in the United States. These 4863 students will be the target group we look at throughout this case study.

Overall, the passing rate of this population is 75.8%. Approximately 45% of this population are first-year students who were admitted into the general engineering program, with the opportunity to declare a Computer Science major in their second year upon passing a set of required courses, including CS1. The majority of the remaining students are science and engineering students, with a large number in our Data Science major. These students require several Computer Science courses as part of their program requirements. Looking at the high school locations, we noticed about 65.7% of students attended a high school in the same state as our institution.

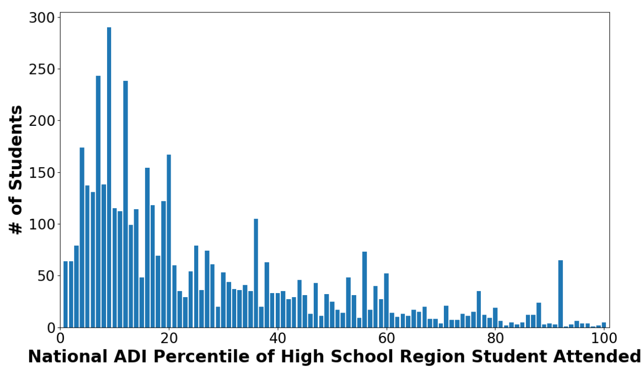

Fig. 4. The population distribution of CS1 students and their corresponding high school region ADI percentile.

After retrieving the neighborhood ADI, census tract SDI and county SDI measure for each student's high school, we noticed that the majority of the deprivation indices were centered around the first quartile for all three measures. The distribution of CS1 students' ADIs can be seen in Figure 4. The SDI census tract and SDI county distributions have a similar trend, with the distribution being heavily skewed to the left. This indicates that the majority of students who took CS1 at our institution in the last seven years came from high schools in more socioeconomically advantaged regions. This was to be expected as our data is collected from a large publicly funded research-based university, with a selective engineering school where the CS department is housed in.

VII. Results

In this section, we will look at socioeconomic percentile splits, inspired by current deprivation measure usage in the medical field [19], [20]. We will also examine the correlations between CS1 students' final grades and their high school region socioeconomic status, which will give us a more fine-grained approach to examining the impact of high school SES on CS1 performance. The sections below describe our results in the context of our research questions.

A. *RQ1: Are students more likely to pass CS1 if they come from a more advantaged high school region socioeconomic status?*

The socioeconomic percentile splits in this case study are 50/50, 85/15, and 15/85. By splitting the student populations into two groups by socioeconomic status, we can compare an advantaged and a disadvantaged group. We first chose a 50/50 split because we wanted to gain a general sense of students attending high school in the top half of advantaged areas vs. disadvantaged socioeconomic high school areas, similar to [19].

We then considered using 10/90 and 90/10 splits, as we wanted to see how students who attended high schools in particularly advantaged or disadvantaged regions performed as inspired by [20]. However, our findings for 10/90 and 90/10 were insignificant due to the small population representation in the bottom 10%. We then expanded the regions of interest to 15/85 and 85/15 to get a larger population on the minority side of the split. Additionally, we also considered splitting the population by quarters and thirds however, we found that the findings in these splits were represented by the splits we had previously chosen.

TABLE I
ADI neighborhood (Census Block Group) percentile splits and the passing rate of CS1 students in these splits

ADI Splits (top %, bottom %)	Passing Rate of Top Split	Passing Rate of Bottom Split	*p-value*
(50, 50)	76.32%	73.23%	**0.038***
(15, 85)	77.57%	74.46%	**0.006****
(85, 15)	75.77%	77.18%	0.344

The passing rates of these splits for ADIs can be seen in Table I. For the ADI census block group measure, we saw that students passed CS1 more often if they were in the more advantaged portions of the 15/85 and 50/50 splits. For both the 15/85 and 50/50 splits, we can see students from the more socioeconomically advantaged groups pass CS1 approximately 3% more often and to a statistically significant degree. Unlike the 15/85 and 50/50 splits, in the 85/15 split, we saw that students in the bottom split of high school area ADIs would pass CS1 more often than students from the top part of the split. However, this finding was not statistically significant as only a small portion of the population was represented in the less advantaged 15% of high school region ADIs. Overall, these results indicate that there is a likelihood to pass CS1 more often if a student previously attended high school in a more advantaged neighborhood.

When looking at the SDI census tract measure, we see very similar trends to the ADI results in the 50/50 and 15/85 splits, as shown in Table II. For these splits, students tend to pass the class between 3% and 4% more frequently in the advantaged regions. Additionally, these results are slightly more statistically significant than the ADI findings. However, unlike the ADI results, in the 85/15 split, we observed students

TABLE II
SDI CENSUS TRACT PERCENTILE SPLITS AND THE PASSING RATE OF CS1 STUDENTS IN THESE SPLITS

SDI CT Splits (top %, bottom %)	Passing Rate of Top Split	Passing Rate of Bottom Split	*p-value*
(50, 50)	76.61%	72.70%	**0.016***
(15, 85)	77.89%	74.22%	**0.003****
(85, 15)	76.00%	72.29%	0.228

pass the class less frequently in the bottom split as compared to the top split. However, again, this split did not provide significant results. The lack of significance of these results may be caused by the broader and more heterogeneous region a county examines. We can similarly observe that students may be more likely to pass CS1 if they are from more advantaged high school regions based on these results.

TABLE III
SDI COUNTY PERCENTILE SPLITS AND THE PASSING RATE OF CS1 STUDENTS IN THESE SPLITS

SDI County Splits (top %, bottom %)	Passing Rate of Top Split	Passing Rate of Bottom Split	*p-value*
(50, 50)	76.34%	73.18%	0.069
(15, 85)	78.37%	74.42%	$<$ **0.001****
(85, 15)	76.06%	70.05%	0.080

Lastly, as represented by Table III, we looked at the largest region deprivation index, the SDI county socioeconomic deprivation measure. While all three splits followed the trend of having the top split pass more frequently, only the 15/85 split proved to be statistically significant. This may be because the two populations on each side of this split are closer to equal than the other two splits, given that the distribution of the deprivation indices is skewed towards the first quartile of advantaged areas. However, for the non-statistically significant 50/50 and 85/15 findings from the SDI county results, we still see the more advantaged split have a higher passing rate on average in both instances.

Based on the results from these three socioeconomic indices, we see that the top 15% of students from more socioeconomically advantaged high school regions pass CS1 more often. Additionally, in the two socioeconomic indices that concern smaller regions surrounding the high school, ADI and the SDI census tract measure, we observe that students from the top 50% of advantaged high school regions pass CS1 more often. Given the consistency of these results, we can conclude that students may be more likely to pass CS1 if they are from a more socioeconomically advantaged high school region.

B. *RQ2: Are students more likely to receive a higher grade in CS1 if they come from a more advantaged high school region socioeconomic status?*

To explore this research question, we first looked to see if there was a correlation between the CS1 students' final grades and their high school SES. Table IV summarizes the result. The lack of correlation between these SES measures and students' final grades indicates that there is not necessarily a correlation between performance and SES measures.

TABLE IV
CORRELATION COEFFICIENTS BETWEEN CS1 GRADE AND SES INDICES

	ADI Percentile	SDI CT Percentile	SDI County Percentile
CS1 Grade	-0.042	-0.033	-0.044

We then looked at the final grade performance groups described in Section 5B ((Excellent, Average, Poor, Fail, Withdraw) and the SES measures. While our below analysis was performed on all three SES measures, we only chose to report our findings about the ADI measure since SDI tract and SDI county analysis produced very similar results.

Figure 5 shows the box plots for the CS1 students' high school neighborhood ADI for each of the performance groups. Most performance groups have a very similar distribution, with the exception of the failing performance group which has a wider distribution and includes more students from the more disadvantaged SES regions.

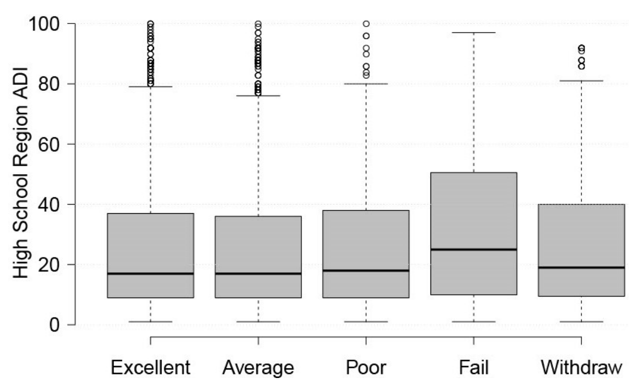

Fig. 5. The CS1 students' high school region ADIs organized by CS1 final grade performance groups.

To analyze these results further, we ran an ANOVA test and saw a significant difference between these performance groups. A post-hoc Tukey HSD analysis was then performed as shown in Table V. The results indicate a statistically significant difference between the failing group and the excellent, average, and poor performance groups. These results show almost no difference between the individual performance groups of students who passed the course. From this data, we cannot conclude that students obtain better grades in CS1 if they are

TABLE V
P-VALUES OF CS1 PERFORMANCE GROUPS' HIGH SCHOOL ADIS USING TUKEY HSD ANALYSIS

Category (*n*)	Excellent	Average	Poor	Fail	Withdraw
Excellent (1393)	—				
Average (2148)	.969	—			
Poor (363)	.999	.987	—		
Fail (227)	**.005***	**.001***	**.038***	—	
Withdraw (539)	.969	.792	.996	.054	—

from a more advantaged high school region. However, this data does imply there may be some difference not between performance groups in CS1, but between the overall passing & failing groups. Additionally, since there is no measured statistical significance between students dropping from CS1 and the other performance categories, we cannot make any conclusions between withdrawing from the course and high school region SES. One hypothesis is that students from lower high school SES may have less support structure and they may not realize that they should withdraw to avoid a failing grade. Further analysis is needed to explore the reasons for the higher failure rates of these populations.

VIII. DISCUSSION

The findings in our study allow us to draw the conclusion that the socioeconomic status of a student's high school most likely has an impact on a student's ability to pass CS1. Throughout all three unique measures used, our result shows that students from regions in the top 15% of nationally advantaged socioeconomic status pass CS1 more frequently and to a statistically significant degree. Additionally, in the two more localized measures, we see that this trend extends to the top 50% of nationally advantaged regions. Our results are in line with previous studies that show that pre-matriculation experiences are a strong indicator of performance, yet we examine this from a socioeconomic standpoint.

However, our results regarding performance groups were inconclusive, similar to [10]. Based on the results of the three socioeconomic indices, we cannot conclude that a student will score higher in CS1 based on high school region SES. For our students, it is very unlikely that coming from a higher high school SES had any effect on their grades. This is promising, as it shows that students who have the knowledge to pass the course can achieve high grades regardless of prior surrounding high school socioeconomic factors.

The methodology and findings in this paper raise the need for further investigation into this method. Possible avenues of future work include examining the high school attributes such as funding and graduation rate, as well as the socioeconomic status of the neighborhood in which the student previously lived. While the socioeconomic status of a region that a high school belongs to does not necessarily reflect the status of an individual student, this is still a good starting point for trying to identify socioeconomic disparities in education.

IX. THREATS TO VALIDITY

A threat to the validity of this paper is that some high schools may not be influenced by the socioeconomic status of the surrounding area. Highly funded secondary schools with an excellent quality of education may exist in a more disadvantaged socioeconomic area, as well as vice-versa. Additionally, another threat to validity may be the lack of information about high schools themselves. The personal socioeconomic status of an individual student and family income may not be reflected in the socioeconomic status of a student's high school. For example, a student from a more disadvantaged area may attend a high school that is in a more socioeconomically advantaged region, which is commonly seen in 'magnet' schools. Lastly, students face a diverse number of reasons for either not completing or failing CS1, such as financial issues, external factors, personal interest in the topic, etc. It is not always the case that because students drop or fail CS1 they were not well-equipped from high school to pass the course.

X. CONCLUSION

This paper introduced a novel method inspired by the current use of socioeconomic deprivation indices in the medical literature. For this study, we use these important measures to examine the relationship between educational success in CS1 and the regional socioeconomic status of a student's high school. Previously, we performed two prior studies that looked at performance indicators using factors such as demographic information, college majors, GPA, prior programming experience, and prior AP credits [5], [31]. However, in those studies, we were not able to use the high school information from the student data we had acquired. This initially inspired us to propose a method to utilize this information and determine whether the socioeconomic status of the high school region affects a student's performance.

Our findings were more nuanced and less intuitive than we had initially expected. Socioeconomic status and computing education is an under-explored research topic. Our goal with this work is to raise awareness for researchers in computer science education to consider a student's socioeconomic status as an important demographic factor. With future work, we hope to further understand if there exists a socioeconomic disparity gap between students and look for interventions to increase equal opportunities of success in CS1.

References

[1] I. Utting, A. E. Tew, M. McCracken, L. Thomas, D. Bouvier, R. Frye, J. Paterson, M. Caspersen, Y. B.-D. Kolikant, J. Sorva, and T. Wilusz, "A fresh look at novice programmers' performance and their teachers' expectations," in *Proceedings of the ITiCSE Working Group Reports Conference on Innovation and Technology in Computer Science Education-Working Group Reports*, ser. ITiCSE -WGR '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 15–32. [Online]. Available: https://doi.org/10.1145/2543882.2543884

[2] S. Bergin and R. Reilly, "Programming: Factors that influence success," in *Proceedings of the 36th SIGCSE Technical Symposium on Computer Science Education*, ser. SIGCSE '05. New York, NY, USA: Association for Computing Machinery, 2005, p. 411–415. [Online]. Available: https://doi.org/10.1145/1047344.1047480

[3] B. C. Wilson and S. Shrock, "Contributing to success in an introductory computer science course: A study of twelve factors," in *Proceedings of the Thirty-Second SIGCSE Technical Symposium on Computer Science Education*, ser. SIGCSE '01. New York, NY, USA: Association for Computing Machinery, 2001, p. 184–188. [Online]. Available: https://doi.org/10.1145/364447.364581

[4] D. Horton and M. Craig, "Drop, fail, pass, continue: Persistence in cs1 and beyond in traditional and inverted delivery," in *Proceedings of the 46th ACM Technical Symposium on Computer Science Education*, ser. SIGCSE '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 235–240. [Online]. Available: https://doi.org/10.1145/2676723.2677273

[5] M. Ellis and S. Hooshangi, "Replication and expansion study on factors influencing student performance in cs2," in *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*, ser. SIGCSE 2023. New York, NY, USA: Association for Computing Machinery, 2023, p. 896–902. [Online]. Available: https://doi.org/10.1145/3545945.3569867

[6] C. K. Jackson, R. C. Johnson, and C. Persico, "The effects of school spending on educational and economic outcomes: Evidence from school finance reforms," National Bureau of Economic Research, Tech. Rep., 2015.

[7] B. D. Baker, "Does money matter in education?" *Albert Shanker Institute*, 2016.

[8] J. A. Thompson, M. Ellis, and S. Hooshangi, "High school socioeconomic neighborhood status and cs1 performance," in *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 2*, ser. SIGCSE 2023. New York, NY, USA: Association for Computing Machinery, 2023, p. 1348. [Online]. Available: https://doi.org/10.1145/3545947.3576294

[9] C. Alvarado, G. Umbelino, and M. Minnes, "The persistent effect of pre-college computing experience on college cs course grades," in *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, ser. SIGCSE '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 876–881. [Online]. Available: https://doi.org/10.1145/3159450.3159508

[10] M. C. Parker, A. Solomon, B. Pritchett, D. A. Illingworth, L. E. Marguilieux, and M. Guzdial, "Socioeconomic status and computer science achievement: Spatial ability as a mediating variable in a novel model of understanding," in *Proceedings of the 2018 ACM Conference on International Computing Education Research*, ser. ICER '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 97–105. [Online]. Available: https://doi.org/10.1145/3230977.3230987

[11] S. R. Sirin, "Socioeconomic status and academic achievement: A meta-analytic review of research," *Review of Educational Research*, vol. 75, no. 3, pp. 417–453, 2005. [Online]. Available: https://doi.org/10.3102/00346543075003417

[12] P. Attainment, "Differences by socioeconomic status," *National Center for Education Statistics*, 2015. [Online]. Available: https://nces.ed.gov/programs/coe/pdf/coe_tva.pdf

[13] A. F. Made and A. Hasan, "Creating a more equitable cs course through peer-tutoring," *J. Comput. Sci. Coll.*, vol. 35, no. 10, p. 33–38, apr 2020.

[14] D. J. Gagnon and M. J. Mattingly, "Limited access to ap courses for students in smaller and more isolated rural school districts," 2015. [Online]. Available: https://dx.doi.org/10.34051/p/2020.227

[15] B. Dalton, E. Hensley, R. Studley, and T. Campbell, "Data and methods for calculating the best high schools rankings: Analytical methodology and technical appendices," 2022. [Online]. Available: https://www.usnews.com/media/best-high-schools/best-high-schools-technical-appendix.pdf

[16] U. News and W. Report, "2022 best u.s. high schools," 2022. [Online]. Available: https://www.usnews.com/education/best-high-schools/national-rankings

[17] B. D. Baker, D. G. Sciarra, and D. Farrie, "Is school funding fair? a national report card," *Education Law Center*, 2014. [Online]. Available: https://eric.ed.gov/?id=ED570455

[18] R. Pampalon, D. Hamel, P. Gamache, A. Simpson, and M. Philibert, "Validation of a deprivation index for public health: a complex exercise illustrated by the quebec index," *Chronic diseases and injuries in Canada*, vol. 34, no. 1, 2014. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/24618377/

[19] J. Mora, A. N. Krepline, M. Aldakkak, K. K. Christians, B. George, W. A. Hall, B. A. Erickson, N. Kulkarni, D. B. Evans, and S. Tsai, "Adjuvant therapy rates and overall survival in patients with localized pancreatic cancer from high area deprivation index neighborhoods," *The American Journal of Surgery*, vol. 222, no. 1, pp. 10–17, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0002961020307777

[20] W. R. Powell, W. R. Buckingham, J. L. Larson, L. Vilen, M. Yu, M. S. Salamat, B. B. Bendlin, R. A. Rissman, and A. J. H. Kind, "Association of Neighborhood-Level Disadvantage With Alzheimer Disease Neuropathology," *JAMA Network Open*, vol. 3, no. 6, pp. e207 559–e207 559, 06 2020. [Online]. Available: https://doi.org/10.1001/jamanetworkopen.2020.7559

[21] R. L. Phillips, W. Liaw, P. Crampton, D. J. Exeter, A. Bazemore, K. D. Vickery, S. Petterson, and M. Carrozza, "How other countries use deprivation indices—and why the united states desperately needs one," *Health Affairs*, vol. 35, no. 11, pp. 1991–1998, 2016. [Online]. Available: https://doi.org/10.1377/hlthaff.2016.0709

[22] D. C. Butler, S. Petterson, R. L. Phillips, and A. W. Bazemore, "Measures of social deprivation that predict health care access and need within a rational area of primary care service delivery," *Health Services Research*, vol. 48, no. 2pt1, pp. 539–559, 2013. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1475-6773.2012.01449.x

[23] A. J. Kind and W. R. Buckingham, "Making neighborhood-disadvantage metrics accessible — the neighborhood atlas," *New England Journal of Medicine*, vol. 378, no. 26, pp. 2456–2458, 2018.

[24] U. of Wisconsin School of Medicine and P. Health, "Area deprivation index," 2022. [Online]. Available: https://www.neighborhoodatlas.medicine.wisc.edu/

[25] A. Documentation, "United states census geography-related concepts: Documentation," 2020. [Online]. Available: https://learn.arcgis.com/en/related-concepts/united-states-census-geography.htm

[26] U. C. Bureau, "Understanding geographic identifiers (geoids)," Oct 2021. [Online]. Available: https://www.census.gov/programs-surveys/geography/guidance/geo-identifiers.html

[27] C. Board, "How to look up ceeb codes," Dec 2022. [Online]. Available: https://blog.collegeboard.org/how-look-ceeb-codes

[28] W. by Liaison "Data dictionaries and master code lists," May 2021. [Online]. Available: https://help.liaisonedu.com/WebAdMIT_Help_Center/Documents_and_Reference_Guides/Data_Dictionaries_and_Master_Code_Lists

[29] "United states census bureau geocoder api," 2022. [Online]. Available: https://geocoding.geo.census.gov/geocoder/

[30] C. Latulipe, A. Rorrer, and B. Long, "Longitudinal data on flipped class effects on performance in cs1 and retention after cs1," in *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, ser. SIGCSE '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 411–416. [Online]. Available: https://doi-org.ezproxy.lib.vt.edu/10.1145/3159450.3159518

[31] S. Hooshangi, M. Ellis, and S. H. Edwards, "Factors influencing student performance and persistence in cs2," in *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education - Volume 1*, ser. SIGCSE 2022. New York, NY, USA: Association for Computing Machinery, 2022, p. 286–292. [Online]. Available: https://doi.org/10.1145/3478431.3499272