Why Diversity, Equity and Inclusion is Still Important

Juan B. Gutiérrez[1*], David Risk-Mora[1] and Paula Jaimes-Buitron[1]

[1*]Department of Mathematics, University of Texas at San Antonio, 1 UTSA Cir, San Antonio, 78249, Texas, USA.

*Corresponding author(s). E-mail(s): juan.gutierrez3@utsa.edu;
Contributing authors: david.risk@utsa.edu;
paula.jaimesbuitron@utsa.edu;

Abstract

Although the benefits of enrolling in Advanced Placement (AP) courses are clear and ideally, every student should have access to and excel in these courses, the reality paints a different picture. Arbitrary factors such as race, ethnicity, ZIP code, and socioeconomic status, collectively known as the opportunity gap, often dictate individuals' educational paths and perpetuate disparities in access to rigorous academic offerings. By utilizing publicly available datasets from the Civil Rights Data Collection (CRDC) and the U.S. Bureau of the Census' Small Area Income and Poverty Estimates (SAIPE) program, alongside specific data from the University of Texas at San Antonio, this research aims to provide a detailed and nuanced understanding of how economic and demographic factors influence educational opportunities. The analysis reveals a significant positive correlation between AP course participation and higher college GPAs, suggesting that students who engage in AP coursework are better prepared for academic challenges in higher education. However, the study finds no statistically significant association between AP participation and college graduation rates. Furthermore, the comparative analysis of the relationship between the proportion of children aged 5-17 living in poverty and enrollment in AP courses in Texas reveals substantial disparities linked to socioeconomic status. Higher poverty rates are associated with lower AP enrollment rates, particularly in math and science. Overall, the findings underscore the critical importance of addressing the opportunity gap to promote educational equity and excellence. By implementing targeted interventions and fostering a supportive and inclusive educational environment, we can ensure that all students, regardless of their socioeconomic background, have the opportunity to achieve their full potential and succeed in their academic and professional pursuits.

Keywords: keyword1, Keyword2, Keyword3, Keyword4

1 Introduction

2 Methods

3 Data Sources

This study utilizes several datasets, including both published and unpublished data. The published data were obtained from publicly available datasets provided by the Civil Rights Data Collection (CRDC), and the U.S. Bureau of the Census's Small Area Income and Poverty Estimates (SAIPE) program. These datasets are openly accessible to researchers and the general public, facilitating transparency and reproducibility in scientific inquiry. No special permissions or data requests were required for access, as the datasets are freely available for download and analysis.

3.1 Civil Rights Data Collection (CRDC)

This study utilizes data from the 2017-18 dataset provided by CRDC. Since 1968, the CRDC, previously known as the Elementary and Secondary School Survey, has gathered information on significant education and civil rights matters within the public school system of the United States [1]. The data obtained from this survey are utilized by various entities, including the U.S. Department of Education's Office for Civil Rights (OCR) for its enforcement and oversight endeavors, as well as by other Department of Education divisions, federal agencies, policymakers, and researchers outside the Department of Education. The CRDC's scope encompasses details concerning school attributes and the provision of programs, services, and outcomes for students. Moreover, the majority of student data is broken down by race/ethnicity, gender, English proficiency, and disability status.

The data collection for the 2017-18 CRDC commenced on January 23rd, 2019, for state educational agencies (SEAs) and on February 4th, 2019, for local educational agencies (LEAs), concluding on June 21st, 2019 [1]. Throughout this period, the CRDC Partner Support Center (PSC) played a vital role in facilitating the data collection process. The PSC provided ongoing assistance to LEAs, addressing both technical and substantive inquiries related to the CRDC. This support encompassed tasks such as logging into the online submission tool, resolving errors, and certifying data, as well as clarifying survey questions and definitions. Additionally, the PSC conducted proactive outreach efforts to ensure engagement from LEAs. This included sending frequent emails containing important dates and tips, conducting weekly follow-ups with uncertified LEAs to offer assistance, and intensifying outreach to unresponsive LEAs. Even after the data collection period had closed, the PSC continued its outreach efforts, requesting data corrections for common errors identified. These activities ensured a comprehensive and accurate data collection process for the

2017-18 CRDC.

Data could be submitted to the online submission tool through two distinct methods: manual data entry or uploading flat files [1]. LEAs were given the flexibility to choose between these methods or utilize a combination of both. Analysis of data submission patterns revealed that a majority of LEAs, comprising 54% of total submissions, opted for exclusive use of the online data entry tool, while the remaining 46% utilized flat file submission for either partial or complete data submission. Furthermore, select SEAs contributed all or part of the data on behalf of their respective LEAs. To support LEAs throughout the submission process, the online tool included a resources page accessible from every screen. This page offered guidance and assistance, including tip sheets on challenging survey items and instructions for avoiding common errors associated with flat file submission. These resources were also available outside of the online tool, providing additional support to LEAs navigating the data submission process.

3.2 Small Area Income and Poverty Estimates (SAIPE)

Data from the SAIPE program, administered by the U.S. Census Bureau, was also used. SAIPE provides annual model-based estimates of income and poverty statistics for states, counties, and school districts [2]. These estimates are crucial for the allocation of federal funds and the administration of both federal and state programs. Specifically, SAIPE offers estimates on the total number of people in poverty, the number of children across various age groups living in poverty, and median household income. For school districts, additional estimates include total population and the number of children ages 5 to 17.

Historically, these estimates on income and poverty statistics for smaller geographic areas such as counties, cities, and school districts were only available from the decennial census long-form, which led to significant gaps in data availability and relevance due to the ten-year interval between censuses [2]. The need for more timely data was recognized in the early 1990s, prompting the formation of a federal agency consortium to fund research for postcensal income and poverty estimates.

In 1993, the U.S. Bureau of the Census initiated the SAIPE program with the support of several federal agencies, including the Departments of Agriculture, Education, Health and Human Services, Housing and Urban Development, and Labor [2]. This initiative aimed to bridge the gap left by the decennial data, especially in light of economic changes that could occur within a decade. The program's development was significantly influenced by legislative actions such as the Improving America's Schools Act of 1994, which mandated the use of updated poverty data for distributing federal educational funds based on the most recent satisfactory data available from the Department of Commerce.

The SAIPE program produces annual estimates of income and poverty for school-age children and the general population at various geographic levels [2]. These estimates are critical for allocating federal funds and assessing local economic conditions. The program's methodologies and outputs are periodically reviewed and evaluated by expert panels, including those from the National Academy of Sciences, ensuring the reliability and accuracy of the estimates for policy-making and research purposes.

The SAIPE program's estimates are not derived directly from enumerations or administrative records, but through a modeling approach that combines data from different sources [2]. This includes survey data, population estimates, and administrative records for counties and states. For school districts, the estimates incorporate model-based county estimates, federal tax information, and multi-year survey data. Key data sources for these models have evolved over time, transitioning from the Annual Social and Economic Supplements of the Current Population Survey to incorporating data from the American Community Survey (ACS) since 2005. This shift has enhanced the accuracy and relevance of the estimates, especially given the ACS's comprehensive coverage of America's changing demographics.

The methodology behind SAIPE's estimates ensures reduced sampling error and improved reliability, making these estimates invaluable for understanding local economic conditions and for educational research, particularly in the assessment of poverty's impact on school funding and children's educational opportunities [2].

The research utilizes the 2017 dataset from the SAIPE program, which provides annual model-based estimates of income and poverty statistics specifically for that year [3]. The 2017 estimates are aligned with the population controls and income concepts used in the American Community Survey (ACS) single-year estimates from the same year. This dataset includes detailed files for each state, the District of Columbia, and the entire United States, containing essential demographics such as the total population, the population of school-age children, and the estimated number of school-age children living in poverty. Each file is identified by the Federal Information Processing Standards (FIPS) state code and includes the Department of Education Common Core of Data (CCD) ID numbers and district names. The school districts covered in the dataset were confirmed via the 2017 school district mapping survey, which reflects the boundaries used during the 2017-2018 school year. This specific focus on 2017 data provides a consistent and comprehensive snapshot of economic conditions affecting school-aged children across various geographical locales, offering a critical basis for analyzing the impact of socioeconomic factors on educational opportunities and outcomes.

3.3 Unpublished Dataset from UTSA

The University of Texas at San Antonio (UTSA) is a public research university located in San Antonio, Texas [4]. Established in 1969, UTSA is a major institution within the

University of Texas System, serving over 35,000 students across multiple campuses. UTSA is designated as a Hispanic Serving Institution (HSI), a recognition that allows it to access federal funding to enhance educational opportunities for Hispanic students. Approximately 59% of UTSA's student body is Hispanic, with over 70% identifying as underrepresented in higher education, and nearly half of its undergraduates are first-generation college students.

An unpublished dataset obtained from the Office of Institutional Research at UTSA was used to analyze the association between AP courses and college outcomes. The dataset includes specific information required for this research that is not available through public databases. The data request number 201910053 was made to the Office of Institutional Research at UTSA. Due to its unpublished status, detailed contents and specific variables cannot be disclosed, but it has been crucial in supplementing the publicly available data and providing comprehensive insights into the research questions.

4 Data Processing

The data obtained from CRDC and SAIPE were imported into a SQL database for analysis. This involved transforming the raw data files into a format compatible with SQL and uploading them into the database using SQL Server Management Studio. The dataset was structured into tables within the SQL database, with each table representing a different aspect of the data. This approach facilitated efficient data management and querying, allowing for complex analyses to be conducted using SQL queries. Prior to analysis, the data were cleaned and preprocessed within the SQL environment to address any missing values or inconsistencies. By utilizing SQL for data storage and analysis, this study benefited from the robust querying capabilities and scalability offered by relational database management systems.

The raw data is systematically organized within the "data" directory, which is hosted on Git for version control and collaborative development. This directory encompasses several subfolders, each serving a distinct purpose:

../data: Potentially serves as a general repository for additional data or as a placeholder for future datasets.

../data/ussd17.xls: Stores data pertinent to the 2017 dataset from the SAIPE program.

../data/2017-18-crdc-data/2017-18 Public-Use Files/Data/SCH/CRDC/CSV: Stores CSV-formatted data from the 2017-18 CRDC Public-Use Files, focusing on school-related information.

../data/2017-18-crdc-data/2017-18 Public-Use Files/Data/SCH/EDFacts/CSV: Stores CSV-formatted data related to schools from the 2017-18 CRDC Public-Use

Files.

../data/2017-18-crdc-data/2017-18 Public-Use Files/Data/LEA/CRDC/CSV: Stores CSV files containing data associated with Local Education Agencies (LEAs) from the 2017-18 CRDC Public-Use Files.

../data/hmda_2017_nationwide_all-records_labels: Stores data from the Home Mortgage Disclosure Act (HMDA) for the year 2017, nationwide, along with corresponding labels.

../data/EDGE_GEOCODE_PUBLICLEA_1718: Stores geographical coding data related to public LEAs for the year 2017-18.

../data/GRF17: Stores data pertinent to GRF17.

Each subfolder houses various CSV or Excel files pertinent to its respective category. These files are processed and converted to CSV format if necessary before being stored as tables in a PostgreSQL database. The names of the tables within the database correspond to the purpose or source of the data, including "GRF17," "CRDC_SCH," "CRDC_SCH_EDFacts," "CRDC_LEA," "HMDA," "GEOCODE," and "ussd17_edited." This systematic structure, coupled with Git version control, ensures efficient data management, collaboration, and reproducibility of the study's findings.

In addition to using SQL for data management and preliminary analysis, Python was employed to further analyze the data extracted from the SQL database. Python scripts were developed to perform more complex data processing and statistical analysis that went beyond SQL's capabilities. These scripts utilized libraries such as pandas for data manipulation and cleaning, numpy for numerical operations, and scipy for more advanced statistical tests. Python's matplotlib and seaborn libraries were also used for generating visualizations that helped in interpreting the trends and patterns within the data. This dual approach of using SQL for structured querying and Python for advanced computations and visualizations leveraged the strengths of both platforms, ensuring a thorough analysis of the datasets. Python's versatility in handling diverse tasks from data cleaning to sophisticated data modeling and visualization significantly enhanced the analytical rigor and depth of the study.

The Python script, "CreateTablesPostgresSQL.py," housed in the "script" folder of the Git repository, serves to automate the setup and preparation of a PostgreSQL database for educational data analysis. Its workflow encompasses several crucial steps: package management ensures environment consistency via a `requirements.txt` file; SQL file execution handles commands on the PostgreSQL database; table definition generation produces Data Definition Language (DDL) statements; data cleaning eliminates empty lines and superfluous spaces in CSV files; database connection facilitates interaction with PostgreSQL. Furthermore, the script iterates through

subfolders within the “data” directory, converting Excel files to CSV, cleansing CSV files, and generating SQL commands for table creation. Requirements include pandas, psycopg2, and openpyxl packages, alongside an environmental variable “PostgreSQL_PWD” for the PostgreSQL password. Additionally, the creation of a “SQL” folder in the work directory is necessary. Ultimately, the script outputs SQL scripts in the “SQL” folder and generates tables in PostgreSQL under the name “CRDB”.

To analyze the relationship between AP course enrollment and poverty levels across school districts, three Python script APAnalyze.py, APMathAnalyze.py, and APScienceAnalyze.py were utilized. Located in the ‘script’ directory, these scripts automate the visual representation and analysis of the relevant data. Initially, it executes several SQL queries to extract data from a PostgreSQL database, specifically focusing on the number of children in poverty and their enrollment rates in general AP, AP Math and AP Science courses. The script then performs advanced data analysis tasks including the generation of polynomial approximations of histograms for the proportion of children in poverty.

For spatial analysis, these three python programs identifies disparities between school districts with high AP enrollment and those with high poverty rates, employing a color-coded system in its visualizations to distinguish between districts, highlighting significant trends and enabling a clear visual representation of the data. This process not only aids in uncovering underlying patterns but also in visualizing educational and economic disparities. The output, including various visualizations, is organized and stored within the ”Figures” directory and its subfolders “APHistograms”, “APMathHistograms”, and “APScienceHistograms”, which are created by the script if they do not already exist.

5 Data Analysis

5.1 Analyzing AP Course Participation

6 Results

7 Discussion

8 Conclusion

Supplementary information. Mention to a github repository

Acknowledgements. There is no funding associated with this research.

Declarations

- Funding: None.
- Conflict of interest/Competing interests: None.

- Ethics approval and consent to participate: Not applicable.
- Consent for publication. Not applicable.
- Data availability: Public data sources as detailed in the methods section.
- Materials availability: Not applicable.
- Code availability: Git hub repository...
- Author contribution: KH wrote the manuscript as her M.Sc. Thesis under the direction of JBG. TV and PP were part of the thesis committee, reviewed the mansucript and provided feedback during preparation of the thesis and submission of the article.

Editorial Policies for:

Springer journals and proceedings: https://www.springer.com/gp/editorial-policies

Nature Portfolio journals: https://www.nature.com/nature-research/editorial-policies

Scientific Reports: https://www.nature.com/srep/journal-policies/editorial-policies

BMC journals: https://www.biomedcentral.com/getpublished/editorial-policies

Appendix A Section title of first appendix

An appendix contains supplementary information that is not an essential part of the text itself but which may be helpful in providing a more comprehensive understanding of the research problem or it is information that is too cumbersome to be included in the body of the paper.

References

[1] Education), O.: 2017-18 CRDC PUBLIC-USE DATA FILE MANUAL. Accessed on March 14, 2024. https://civilrightsdata.ed.gov/assets/for-researchers/user-manual/2017-18%20CRDC%20Public-Use%20Data%20File%20Manual.pdf

[2] Bureau, U.S.C.: ABOUT THE SMALL AREA INCOME AND POVERTY ESTIMATES (SAIPE) PROGRAM. Accessed on March 14, 2024. https://www.census.gov/programs-surveys/saipe/about.html

[3] Bureau, U.S.C.: 2017 SMALL AREA INCOME AND POVERTY ESTIMATES (SAIPE) FOR SCHOOL DISTRICTS. Accessed on March 14, 2024. https://www.census.gov/data/datasets/2017/demo/saipe/2017-school-districts.html

[4] San Antonio, U.: About UTSA. Accessed on March 14, 2024. https://www.utsa.edu/about/