Title I and Racial Demographic Shifts in U.S. Public Schools

HarvardX PH125.9x Capstone Project - Part 2

Elan Weingarten

2025-06-11

Contents

1 Introduction

Topic: Are Title I schools more likely to see increases in % White students and decreases in % Black or Hispanic students?

This study investigates whether Title I status is associated with different patterns of racial demographic change at the school level over time. Specifically, we analyze whether Title I schools are disproportionately likely to experience decreases in the percentage of Black and Hispanic students and increases in the percentage of White students.

Using both linear and mixed-effects regression models, we aim to capture and predict these trends in racial composition. Our goal is to inform education policy and the public as a whole by patterns in how financial-equity driven funding in public schools is related to the demographic shifts occurring in individual schools.

Background

Title I funding is a provision of the Elementary and Secondary Education Act that gives federal funds to schools with high percentages of students from low-income families. It is widely used as a proxy for school socioeconomic disadvantage. Whether or not a school has a Title I program is based its students' families economic need, and thus analyzing demographic shifts within Title I schools offers insights into broader trends of racial and socioeconomic changes and related government support (NCES Fast Facts: Title I).

Race is used in this analysis because of its centrality to historical and contemporary educational inequities, as well as its importance to the efficacy of school educators (Dee, 2004). Understanding racial composition trends in schools is essential for evaluating the effectiveness of educational policy and to inform future choices regarding public school policy and funding.

---

2 Methods

2.1 Overview

We gathered data from the **National Center for Education Statistics (NCES)** using their **Elementary and Secondary Information System (ELSi)** web tool. Using ELSi, we created and downloaded custom tables containing yearly school-level data, including total enrollment, Title I status, and racial composition among many other variables.

Each row in the dataset represents a single school in a single school year.

To prepare the data for analysis:

- Race and gender counts were transformed into percentages of total enrollment.
- Binary markers (e.g., Title I status, charter school designation) were expanded into columns coded as 1 (yes) or -1 (no).
- School years were normalized to a single year format: for instance, the 2023–2024 school year was labeled as 2024.

We removed data rows from the following:

- **Schools with fewer than 50 students:** These schools represent statistical outliers and may produce unstable estimates due to small sample sizes.
- **Schools missing demographic or Title I data.**
- **Schools with only 1 year of data**

The cleaned dataset included approximately 837,000 rows and 36 columns, representing 93,729 unique schools across 16,954 in all 50 states and the District of Columbia.

Of these, approximately 537,000 had complete data on both racial demographics and Title I status.

2.2 Modeling Approach

Our analysis followed the following structure:

1. Load, clean, and process the dataset
2. Exploratory data analysis
3. Attempt basic linear regression model
4. Refine the model using mixed-effects models
5. Attempt to make predictions using the mixed-effects models
6. Explain the trends

2.3 Load, clean, and process the dataset

```r
# Load data generated by the .R script `Capstone - Education Data Weingarten.R`
load(file.path("data", "r_variables.RData"))

kable(
  as.data.frame(t(dim(ed_data_full))) %>%
    `colnames<-`(c("Rows", "Columns")),
  caption = "Clean data was created from NCES dataset with the following dimensions"
)
```

Table 1: Clean data was created from NCES dataset with the following dimensions

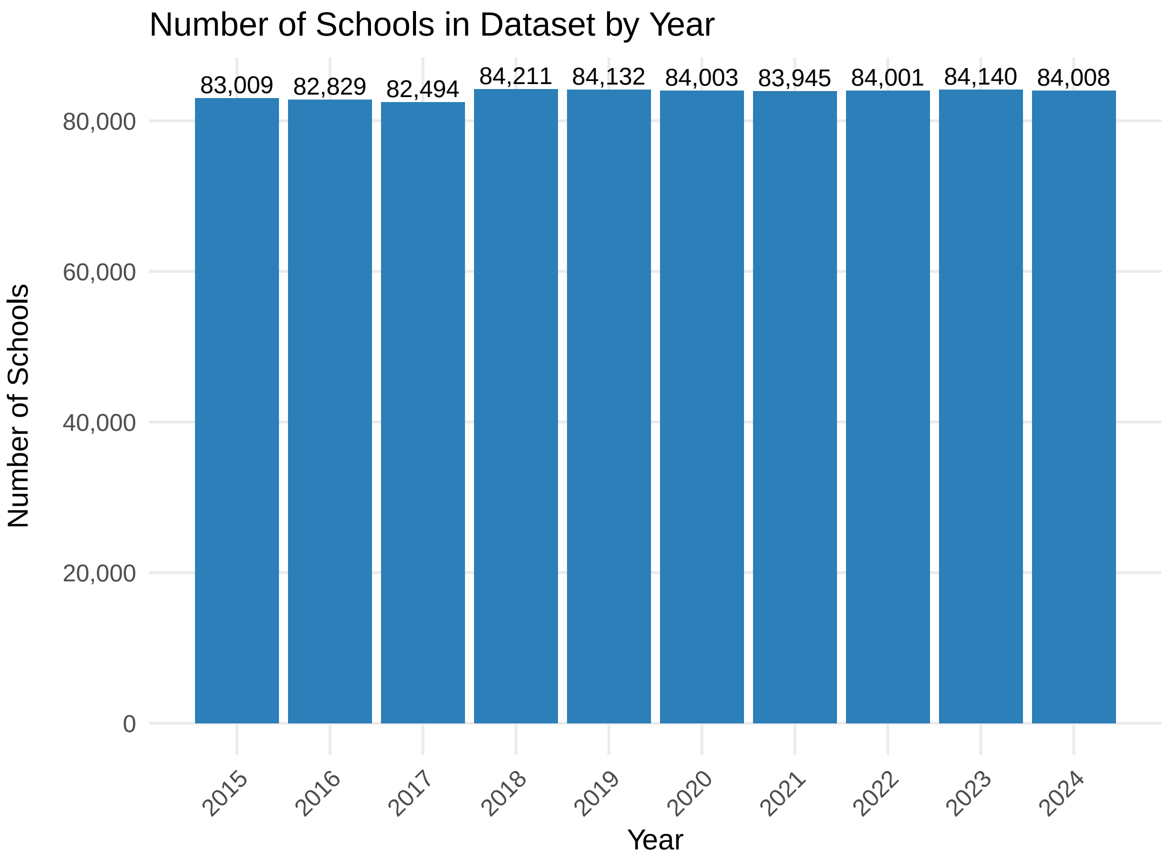
Rows	Columns
836772	36

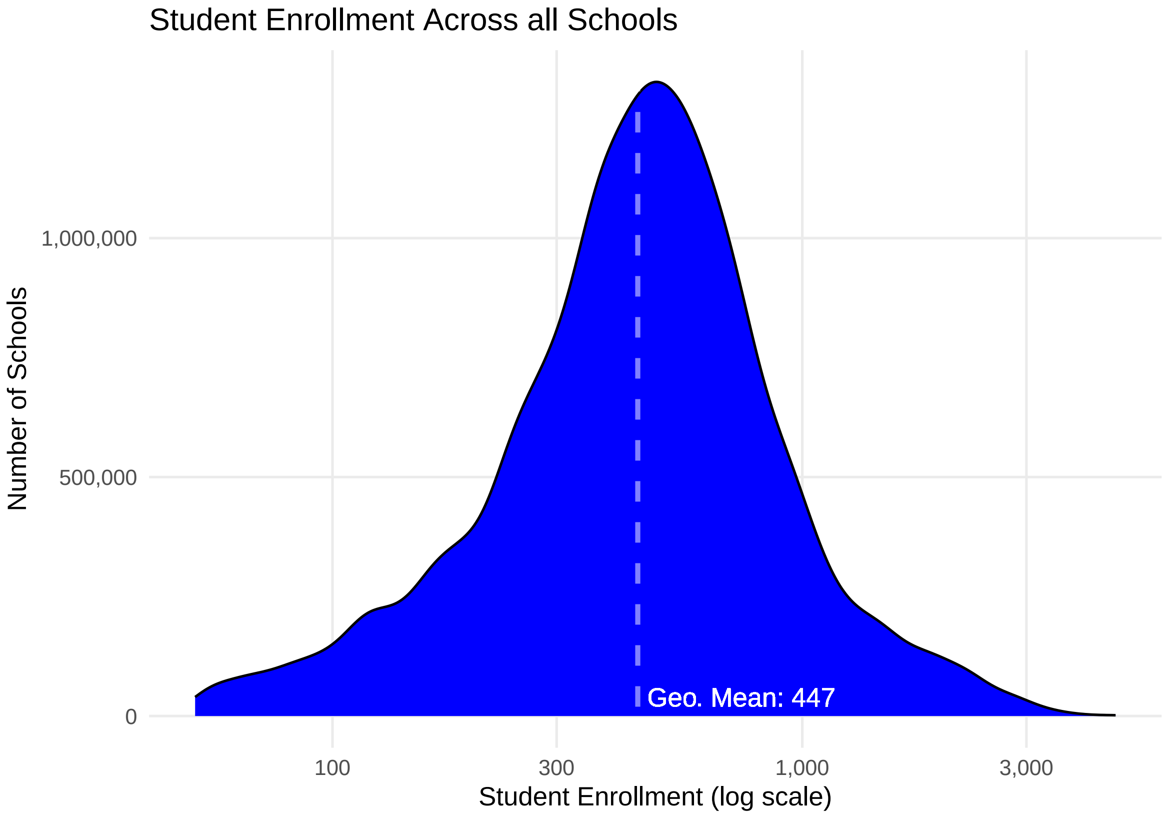
2.4 Exploratory Data Analysis

2.4.1 Counts of schools, districts, and students

The full dataset contains on average roughly 15,100 districts and 83,700 schools per each year of data.

Enrollment across schools is normally distributed (on the logarithmic scale) as to be expected with population sizes. The typical school has 447 students (geometric mean).


Number of Schools in Dataset by Year
83,009
82,829
82,494
84,211
84,132
84,003
83,945
84,001
84,140
84,008
80,000
60,000
40,000
20,000
0
Number of Schools
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
Year

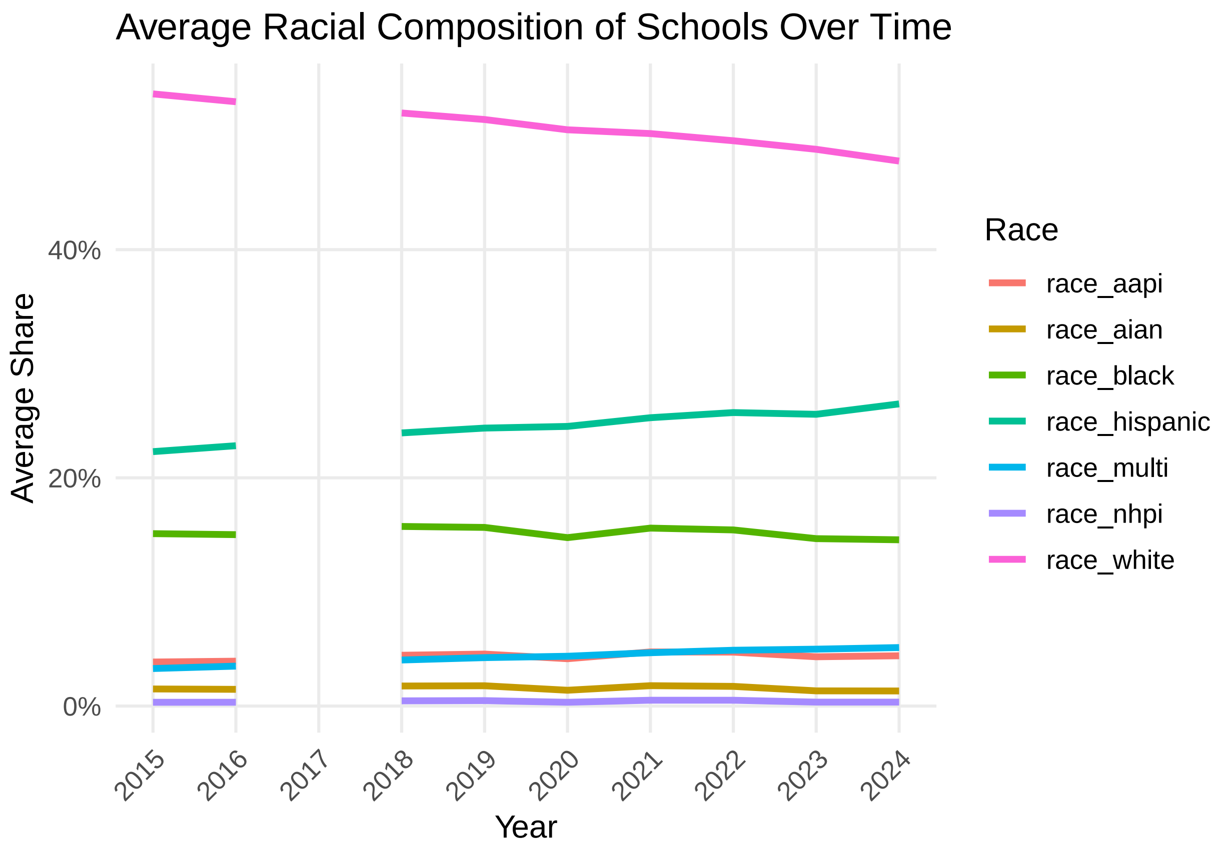

Student Enrollment Across all Schools
1,000,000
500,000
0
Number of Schools
Geo. Mean: 447
100
300
1,000
3,000
Student Enrollment (log scale)

2.4.2 Student racial trends

There is no racial data for 2017, but otherwise a clear trend is shown for all groups from 2015 to 2024:

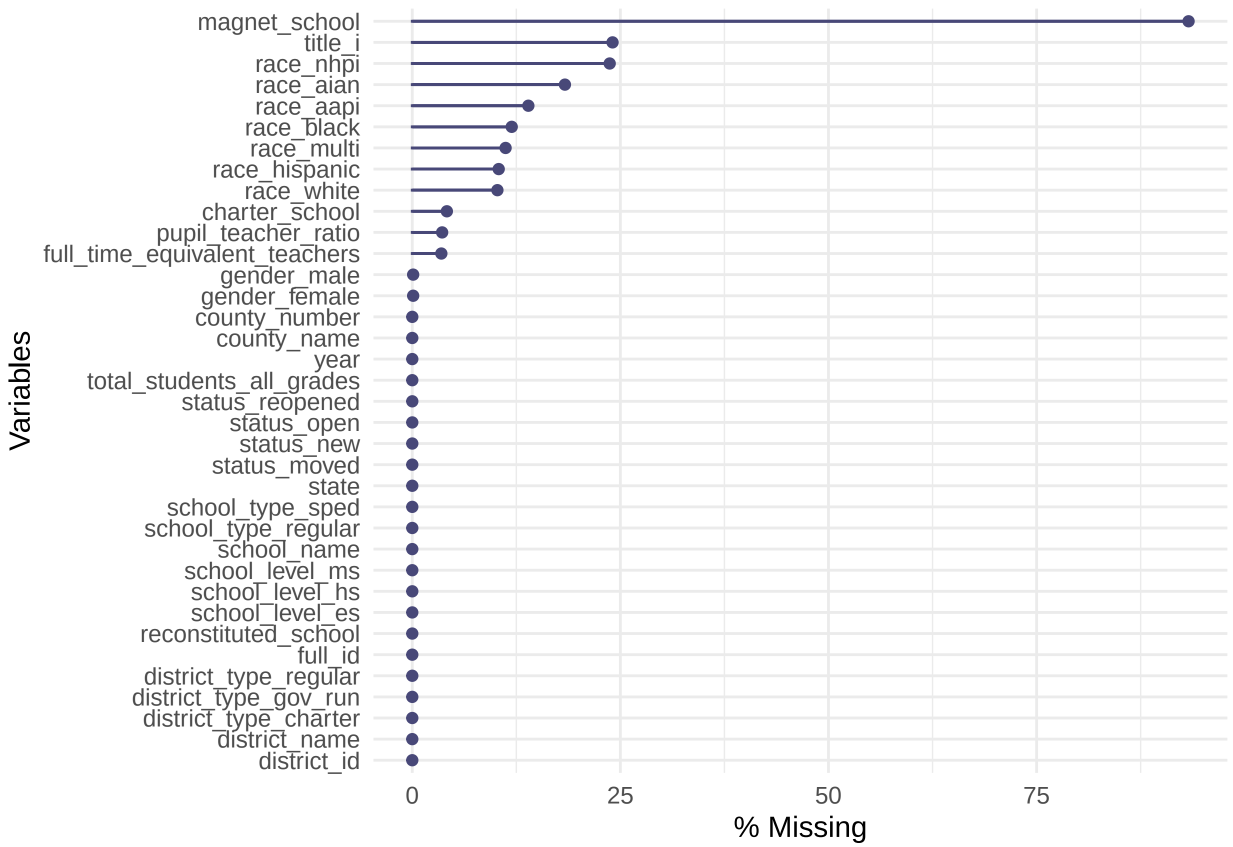
- White is largest group, decreasing steadily.
- Hispanic is second most common group, increasing faster than any other group.
- Black is the third most common group, fluctuating, but steady over all.

All other racial groups form a small minority compared to these three.



2.4.3 Explore missing data

We found that data about the status of a school being a Magnet School data is missing more than 80% of the time, and was thus dropped.



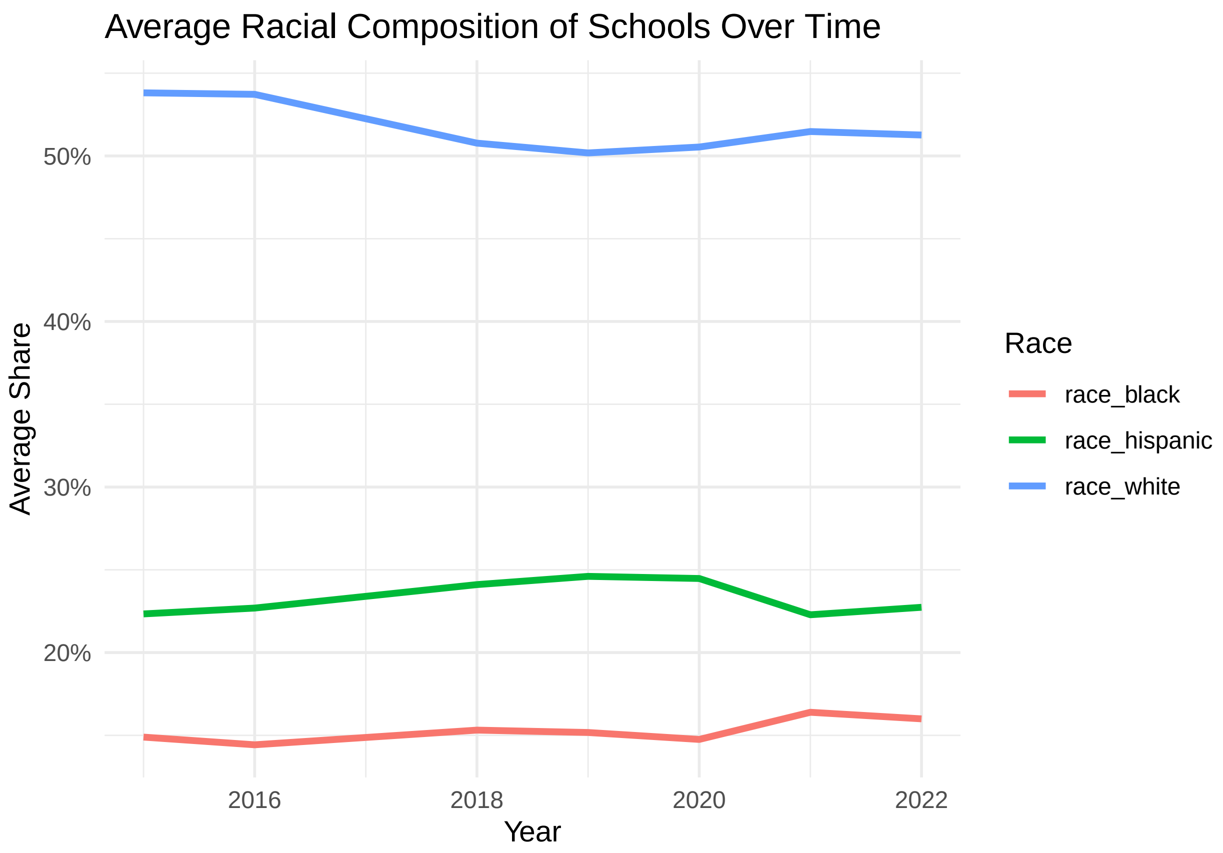
2.4.4 Refocus the dataset

Based on patterns observed during exploratory data analysis, we chose to focus on racial demographic trends in relation to whether or not a school was implementing a Title I program (a proxy for the school educating many students from low-income families). Because income equity efforts can intersect with racial equity concerns, this analysis explores whether Title I schools exhibit different racial trends over time.

To keep the analysis both meaningful and manageable, we focused on the three largest racial groups in the dataset: White, Hispanic, and Black.

2.4.5 Explore racial trends of the top three most populous races

When we focused on these three races, we found that the trend for all races seem to have shifted around the time COVID was having the biggest impact in schools: 2020~2021. Noticing that complication, we chose to investigate whether general trends were still meaningful related to race changes over the years.

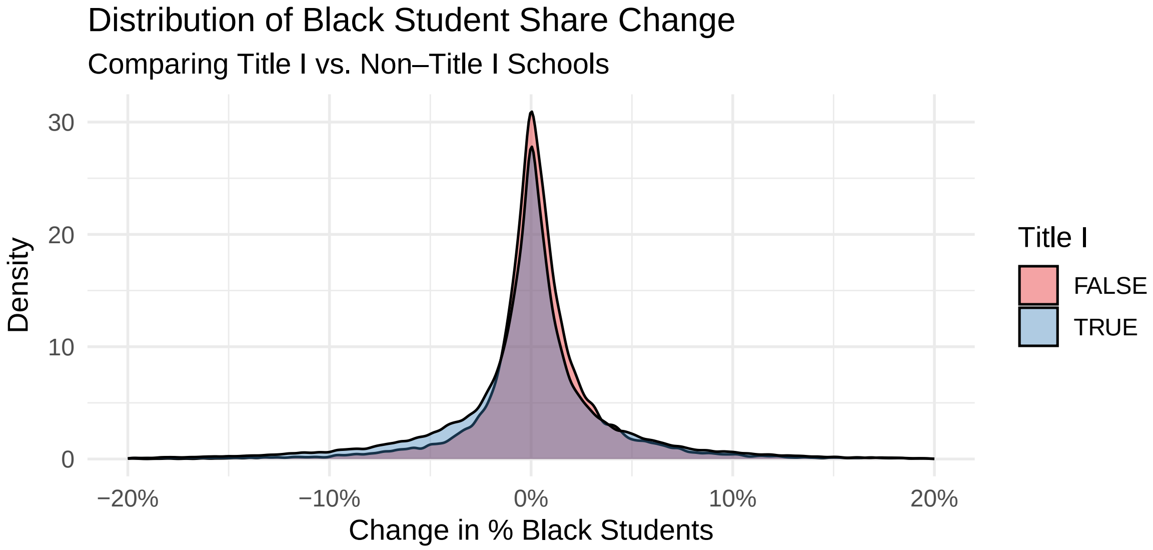


2.4.6 Explore trends in changes in racial shares over the years

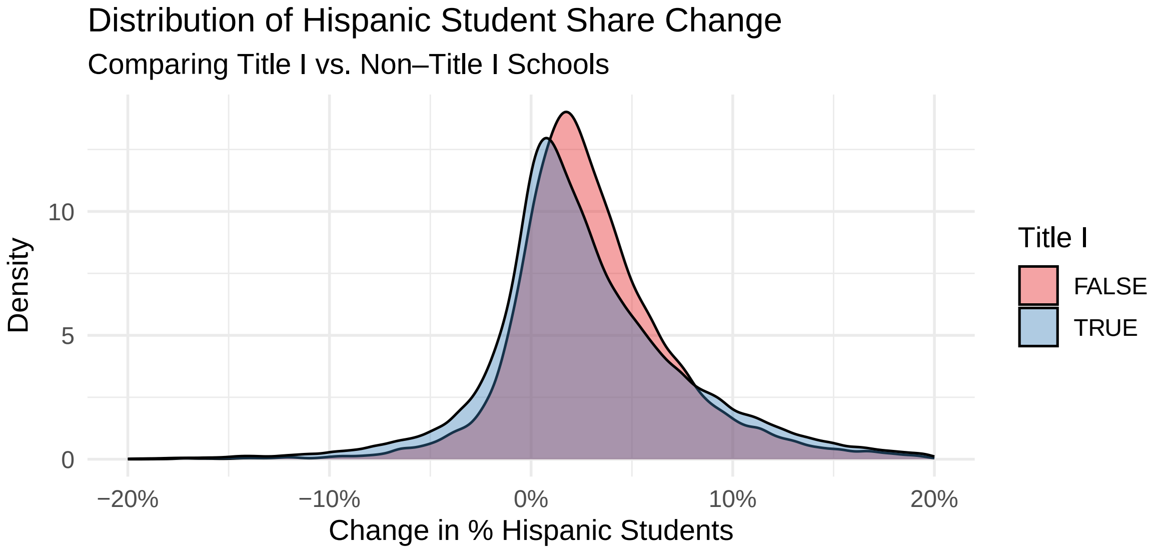
We calculated the change in racial shares by measuring the share of that race for the school's first year in the dataset, and then in its last year of the dataset. By subtracting the last year from the first year's race share, we found the change in racial share.

By plotting the density curves of each change in racial share for both Title I and non-Title I schools we found that:

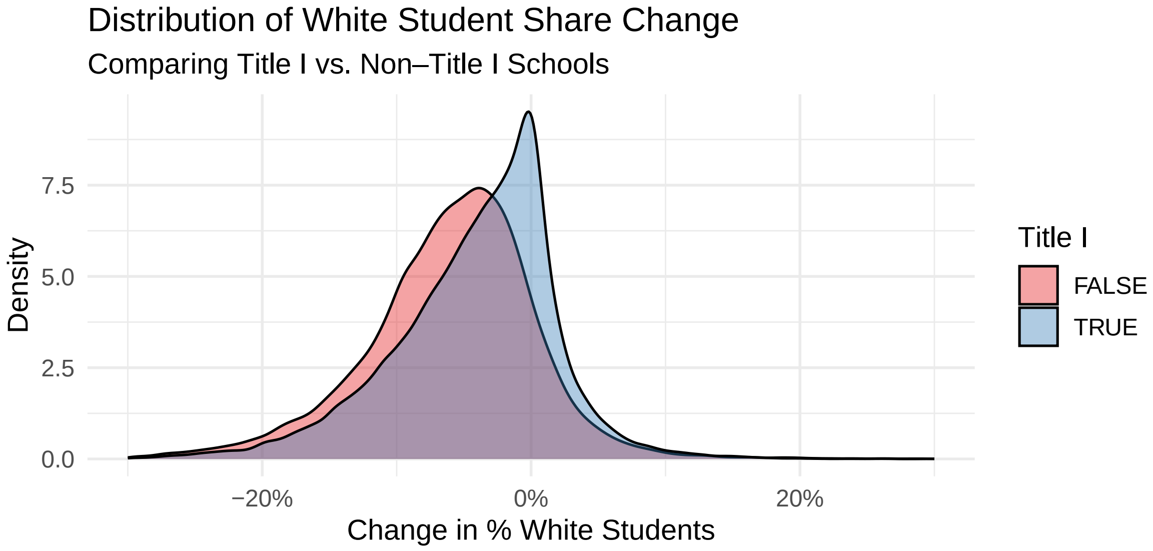
- Title I schools appear to be **more** likely to have lost Black students proportionally in the last 10 years
- Title I schools appear to be **more** likely to have lost Hispanic students proportionally in the last 10 years, more so than Black students.
- Title I schools appear to be **LESS** likely to have lost White students than non-Title I schools in the last 10 years.


Distribution of Black Student Share Change

Comparing Title I vs. Non–Title I Schools


Distribution of Hispanic Student Share Change

Comparing Title I vs. Non–Title I Schools


Distribution of White Student Share Change

Comparing Title I vs. Non–Title I Schools

Table 2: Linear Regression Estimates (Standard Error) [t-values]

Predictor	% Black	% White	% Hispanic
(Intercept)	0.00343 (0.00049) [6.99]	-0.04654 (0.00066) [-70.54]	0.01987 (0.00055) [36.01]
title_i_mode	-0.00586 (0.00044) [-13.33]	0.01598 (0.00059) [27.09]	0.00041 (0.00049) [0.84]
avg_enrollment	0.00000 (0.00000) [2.04]	-0.00002 (0.00000) [-33.72]	0.00002 (0.00000) [33.57]

3 Results

In our exploratory analysis we found an apparent, and surprising, relationship between racial composition and Title I status. To investigate this trend more rigorously, we applied regression modeling techniques to quantify and test these patterns.

3.1 Linear regression model

We began by fitting a basic linear regression model using the `lm()` function. This model aimed to describe and predict the **change in racial composition** (i.e., year-over-year change in % Black, % Hispanic, and % White students) at the school level in relation to Title I status.

Two predictor variables were used:

- **`title_i_mode`**: the most frequent Title I status of a school across the years of data available (0 or 1)
- **`avg_enrollment`**: the school's average enrollment across those years

The results in the table 2, above, show the effect of these two predictor variables on the racial share for each of the three racial groups. It also presents the **standard errors** (in parentheses) and **t-values** (in brackets).

A |t-value| greater than 2 indicates statistical significance at the 95% confidence level. In this model, the Title I coefficient for Hispanic student change is not statistically significant, as its t-value is below this threshold, but all other coefficients were found to be statistically significant.

Note that for this model, the 2014-15 school year was defined as 'year 0', thus the intercept refers to 2015.

```
lm_black = lm(race_black_change ~ title_i_mode + avg_enrollment,
              data = race_trends)
lm_hispanic = lm(race_hispanic_change ~ title_i_mode + avg_enrollment,
                 data = race_trends)
lm_white = lm(race_white_change ~ title_i_mode + avg_enrollment,
              data = race_trends)
```

3.2 Refine the model using mixed-effects regression

Because we found a significant correlation between most of our independent variables and race changes, we investigated a more robust model using Mixed Effect linear regression and the `lmer` function. This models the school (`full_id`) and state as random effects to control for unobserved, school- or state-specific characteristics.

Our first attempt controlling for both the school and the state failed, as the model could not converge for the model for Black or White students. Thus, we tried again, with only the school (`full_id`) as a random effect, since the school already accounted for location and including state was somewhat redundant.

Unlike the earlier model that predicted year-over-year *change*, these models analyze **absolute racial composition** by year, enabling a more precise and stable estimation.

The results below show mixed-effect estimates, along with the standard error (in parentheses) and the t-values (in brackets). Trends show that while Title I schools are losing black and Hispanic students and gaining white students, Non-Title I schools are diversifying

Table 3: Fixed Effects Estimates (Standard Error) [t-values]

Predictor	% Black	% White	% Hispanic
(Intercept)	0.14927 (0.00081) [184.58]	0.54367 (0.00113) [482.61]	0.22217 (0.00092) [242.14]
title_i	0.00624 (0.00018) [33.78]	-0.01637 (0.00025) [-64.67]	0.00509 (0.00022) [23.31]
year_scaled	0.00055 (0.00003) [20.51]	-0.00883 (0.00004) [-241.70]	0.00451 (0.00003) [143.03]
title_i:year_scaled	-0.00091 (0.00003) [-29.63]	0.00311 (0.00004) [73.70]	-0.00043 (0.00004) [-11.92]

All trends with Mixed Effect model are highly statistically significant given the high t-statistic value. Because they are such robust trends, the small magnitudes (ranging from 0.001% to 0.02% shifts in racial share per year/Title I status change) of the effects are not negligible. While these impacts may be small over one or two years at an individual school, when thousands or millions or students on a district or state level are considered, these effects can represent many students.

```
# lmer_black_state = lmer(race_black ~ title_i * year_scaled +
#                           (1 | full_id) + (1 | state), data = ed_data)
# lmer_hispanic_state = lmer(race_hispanic ~ title_i * year_scaled +
#                              (1 | full_id) + (1 | state), data = ed_data)
# lmer_white_state = lmer(race_white ~ title_i * year_scaled +
#                           (1 | full_id) + (1 | state), data = ed_data)
#
# # Error! Model could not converge for black or white.

lmer_black = lmer(race_black ~ title_i * year_scaled +
                    (1 | full_id), data = ed_data)
lmer_hispanic = lmer(race_hispanic ~ title_i * year_scaled +
                       (1 | full_id), data = ed_data)
lmer_white = lmer(race_white ~ title_i * year_scaled +
                    (1 | full_id), data = ed_data)
```

3.3 Attempt to make predictions using the mixed-effects models

The `lmer` model proved too memory-intensive to reliably generate predictions. As a fallback, we used a simpler linear regression model (`lm`) to estimate trends.

We conducted this test by splitting the data in a training (80%) and test (20%) set. We trained the linear regression model on the training set and predicted values for the test set, computing a Root Mean Square Error (RMSE) to determine the accuracy of the model's predictions.

Although the model is trained with highly precise estimates, the RMSE ended up being far to close to the order of magnitude of the demographic percentages (10~50%) that we were already looking at.

```
set.seed(42)  # for reproducibility

# Split ed_data into train (80%) and test (20%)
train_indices <- sample(nrow(ed_data), size = 0.8 * nrow(ed_data))
ed_train <- ed_data[train_indices, ]
ed_test <- ed_data[-train_indices, ]

# Train simpler lm model
simple_lm_black = lm(race_black ~ title_i * year_scaled, data = ed_train)
simple_lm_hispanic = lm(race_hispanic ~ title_i * year_scaled, data = ed_train)
simple_lm_white = lm(race_white ~ title_i * year_scaled, data = ed_train)
```

```r
# Predict on test data
ed_test = ed_test %>%
  mutate(
    pred_black = predict(simple_lm_black, newdata = ed_test),
    pred_white = predict(simple_lm_white, newdata = ed_test),
    pred_hispanic = predict(simple_lm_hispanic, newdata = ed_test)
  )

# Calculate RMSE
rmse_black = RMSE(ed_test$pred_black, ed_test$race_black)
rmse_white = RMSE(ed_test$pred_white, ed_test$race_white)
rmse_hispanic = RMSE(ed_test$pred_hispanic, ed_test$race_hispanic)

# Print results
kable(
  data.frame(
    Race = c("Black", "White", "Hispanic"),
    RMSE = c(rmse_black, rmse_white, rmse_hispanic)
  ),
  digits = 3,
  caption = "Root Mean Squared Error (RMSE) by Race",
  col.names = c("Race", "RMSE")
)
```

Table 4: Root Mean Squared Error (RMSE) by Race

Race	RMSE
Black	0.229
White	0.316
Hispanic	0.255

4 Conclusion

The unreliable prediction results from the linear regression model suggests that while the overall trends may be statistically significant, the linear model lacks the complexity to capture meaningful patterns at the prediction level. More sophisticated approaches — or a richer dataset — may be required to generate reliable forecasts.

Despite those shortcomings, we discovered a surprising trend. Though small, there is indeed a statistically significant trend that **Title I** schools are **losing Black and Hispanic** shares over time more quickly (or gaining them more slowly) than **non–Title I** schools. Conversely, **White** student share is **decreasing** less rapidly in **Title I** schools than in **non–Title I** schools.

There is a clear trend across schools that as the American population grows more diverse, the portion of white students in schools decreases. Some interesting implications of this include bucking a trending idea that it is only poor students of color who benefit from Title I funding or other financial assistance. In fact, despite the trend that white students are decreasingly the majority in public schools, they are proportionally being more represented in schools with Title I programs when compared to their Black and Hispanic counterparts. This may be due economic changes among different races, or in the way that Title I programs are being implemented.

To further understand the trends we are seeing, a more robust modeling approach—beyond linear and mixed-effects regression—would be needed, and other variables would need to be included to control for more variation. However, this report serves as a strong motivating first step in looking further into this trend and understanding what may be driving it, and what the implications may be. With those things understood, perhaps a better predictive model could be created that could aid schools in predicting their racial and economic demographics.

5 References & Acknowledgments

- National Center for Education Statistics (NCES) Elementary/Secondary Information System (ELSi). https://nces.ed.gov/ccd/elsi/
- NCES Fast Facts: Title I https://nces.ed.gov/fastfacts/display.asp?id=158
- Dee, Thomas. (2004). Teachers, Race, and Student Achievement in a Randomized Experiment. The Review of Economics and Statistics. 86. 195-210. 10.1162/003465304323023750.
- Open AI's ChatGPT for code optimization