An Investigation Into Race and Education

Laura Bryant, Darian Shane Martos

December 3, 2021

Abstract

The goal of this paper is to identify the causal effect of race on the educational attainment of Americans from generation to generation (i.e. intergenerational education mobility). We first reconstructed an analysis conducted by John Ferrare as shown in *Intergenerational Education Mobility Trends by Race and Gender in the United States*. Second, we conducted a causal analysis. From results, we find that race still has a big effect on education mobility when accounting for past parents' educational attainment, while generational differences in mobility and race are not as present when accounting for other covariates such as income or family income.

Introduction

In the United States higher education is key to improving our quality of life via "income, cultural capital, [and] social ties" (Ferrare, 2016). As a result, identifying the relationships between a variety of covariates and the difference in educational attainment from generation to generation (hereby referred to as intergenerational education mobility [1]) is a hot topic for many researchers. Specifically, we find the influence of race to be particularly interesting. Given the past and present racial discrimination in the United States, exploring the affect of race on this topic is imperative.

Our analysis is based on the article *Intergenerational Education Mobility Trends by Race and Gender in the United States* by John Ferrare. Ferrare's analysis aimed to identify whether the patterns of intergenerational education mobility differed between white men, white women, black men, and black women. Furthermore, he sought to understand whether these patterns varied with respect to a parent's education levels. The first section of our experiment replicates the ordinary least squares regression Ferrare conducted [2] with a few significant differences. First, we used the latest GSS data which includes data from 2018 and 2021. Ferrare only used the data available as of 2016. Second, we included Hispanic respondents in our analysis because there is evidence that a difference in educational attainment based on race and Hispanic ethnicity exists (Crissey, 2007). The second section of our analysis focuses on identifying the causal effect of race on intergenerational educational mobility whereas Ferrare's focused on identifying a correlation between race and gender on the outcome.

1 Dataset and Handling Imputed Values

We used survey data from the GSS 1972 - 2021 cumulative data file (Davern et al., 2021) that is conducted by the National Opinion Research Center at The University of Chicago. The data is available for public use on https://gss.norc.org/get-the-data. Beginning in 1972, this data was collected via cross-sectional surveys annually until 1990, after which it was conducted every other year[3]. The goal of the GSS is to track trends in public opinion, attitudes, and behaviors. The items included in the interview are comprised of a set that occurs every iteration and a set of topical modules that will change each time.

Our initial dataset had 68,846 observations of 6,311 variables. We created a subset to use for our analysis based on a few filters. First, we dropped all respondents that did not have any education educational attainment data ("educ"). Second, we removed all respondents older than 65. Including them may bias education mobility upwards since there is a correlation between education and life expectancy (Krueger, Tran, Hummer, and Chang, 2015; Olshansky et al.,

[1] This term is coined by Ferrare.
[2] We are referring to the regression whose results are displaying in Table 3.
[3] Due to the pandemic, the survey was conducted in 2021 instead of 2020

2012). Third, we created a subset of all observations using only the 10 variables of interest for our analysis [4].

After apply these filters, we were left with 50,184 observations. Due to the variation between interview items each time the survey is conducted, there are many missing values. Rather than dropping any observations, we decided to employ multiple imputation using the predictive mean matching method to estimate those values.

2 Analysis and Results

Linear Model Re-Analysis

The first part of this project our replicating the analysis of Ferrare, namely fitting linear models across units that do not have missing data. The intent of doing a "re-analysis" of the project is to understand differences of race for data over later cohorts.

Similar to Ferrare, we use the following covariates:

- **race** - In the GSS file itself it is actually a tri-nary variable with (1 = White, 2 = Black, 3 = Other). For the sake of simplicity and to utilize methods from this course, we binarize the race variable to be White = 0, and Black and Other = 1. This helps discern effects between people of color and white individuals in their educational mobility patterns.
- **sex** - Binary indicator variable for respondent sex.
- **educ** - Respondent's highest year of education attained.
- **maeduc/paeduc** - Mother or father education level, this is used for two purposes (1) to model a **pared** parental education variable that corresponds to the higher-attaining parent's years of education and (2) to model our response **mobility**, which is defined as:

$$\textbf{mobility} = \textbf{educ} - \textbf{pared}$$

- **paocc10** - The father's current occupation.
- **sibs** - The number of the respondent's brothers and sisters (biological or adopted). This includes step-siblings and siblings born alive who are now deceased.
- **incom16** - The respondent's best guess at their family's income level compared to other American's when they were 16. The values are one of the following: far below average, below average, average, above average, and far above average.

[4] The 10 variables are maeduc, paeduc, educ, paocc10, sibs, incom16, family16, race, sex, cohort

- **family16** - Whether the respondent was living with their mother and father when they were 16.
- **cohort** - The respondent's birth cohort.

Using the non-imputed data (i.e. all rows/respondents that do not contain missing data), we follow the work of Ferrere and construct four separate linear models that fit regress over a variety of covariates and interactions to predict educational mobility. The table below summarizes estimates for these linear fits. All of the covariates and interactions used are on the left-hand side of the table. The associated models and coefficient estimates and standard errors indicate what covariates serve as features for the corresponding model From the below re-analysis table with coefficient estimates, we take note of a couple of points:

- Regressing over the Cohort^2 covariate provides no added benefit to mobility. Though Ferrare tested over a variety of interactions terms with cohort and squared-cohort to control for differences in mobility across cohorts of respondents, the various models indicate that the time period of the respondents has no discernible effect on mobility for future generations. This is investigated further later in this project.
- In fact, Model 4 shows the drawback of employing too many interactions and squared covariates. Many of the coefficient estimates in Model 4 explode along with their standard error, while many other variables "zero out" and have minimum to no effect on educational mobility.
- Father's occupation also has no significant predictive power towards mobility, regardless of model.
- The main focus of our analysis, the race variable (known as "Black" in the original paper) has varying predictive power over mobility depending on the other covariates and interactions present. Notably, the race variable indicates that being Black or Other (i.e. a "one unit increase in the race variable") leads to a general decrease in educational mobility.

The last bullet point is what we focus on within our causal analysis. We had also ran models akin to Appendices 1 and 2 of the original paper, building linear and logistic regression fits to investigate terciles, which we approach in depth in the subsequent section. We defer that analysis to staff for questions due to the space constraints of this paper. The resulting analysis is interesting though in investigating parental education of respondents and how it interacts with other covariates.

Re-Analysis of Ferrere Table 3

Ordinary Least Squares Regression of Educational Mobility on Race, Gender, Birth Cohort, and Family Background (N = 26,095)

	Model 1		Model 2		Model 3		Model 4	
	B	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*
(Constant)	1.722	0.073	9.926	0.103	-0.004	0.055	-6979	1867
Black	-0.251	0.192	-0.642	0.144	-0.002	0.001	-1304	3639
Male	-0.166	0.046	-0.212	0.034	0.001	0.004	3443	1152
Black × Male	0.503	0.116	0.511	0.087	0.302	0.087	-1889	2248
Parental education			-0.643	0.005	-0.668	0.005	248	144
Father's occupation			≈ 0	≈ 0	≈ 0	≈ 0	≈ 0	≈ 0
Single Mother			-0.071	0.015	-0.069	0.015	-0.069	0.016
Siblings			-0.162	0.006	-0.163	0.006	-0.163	0.005
Income			0.17	0.02	0.181	0.02	0.18	0.02
Cohort					3.871	0.561	7.201	1.919
Cohort^2					0.001	≈ 0	≈ 0	≈ 0
Black × Cohort					3.337	1.461	1.241	3.734
Black × Cohort^2					0.001	≈ 0	≈ 0	≈ 0
Male × Cohort					-1.434	0.348	-3.556	1.184
Male × Cohort^2					≈ 0	≈ 0	≈ 0	≈ 0
Black × Male × Cohort					-0.017	0.893	1.979	2.306
Black × Male × Cohort^2					≈ 0	≈ 0	≈ 0	≈ 0
Parental Education × Cohort							-0.256	0.148
Parental Education × Cohort^2							≈ 0	≈ 0
Parental Education × Black × Cohort							-0.062	0.318
Parental Education × Black × Cohort^2							≈ 0	≈ 0
Parental Education × Male							-0.016	0.895
Parental Education × Male × Cohort							0.169	0.092
Parental Education × Male × Cohort^2							≈ 0	≈ 0
Parental Education × Black × Male × Cohort							-0.048	0.198
Parental Education × Black × Male × Cohort^2							≈ 0	≈ 0
Adjusted R^2	0.004		0.446		0.4558		0.4565	

The table varies greatly from the original version presented in Ferrere's work. The above points address these differences, but the most prominent for our investigation is the coefficient estimate explosion within Model 4. With these estimates being in the thousands, we see that simplifying the analysis by binarizing the race variable and interacting it with cohort gives larger estimates than we had initially expected.

Regardless of the estimates themselves, there is at least some indication of a trend of cohort, which very broadly measures generational differences in educational attainment, moving mobility upwards over the years. The coefficient estimate of cohort asserts this within both models 3 and 4. Though cohort indicates some positive relationship with mobility, we wondered if this upwards trend in mobility applies to racial differences of respondents as well.

Causal Analysis - An Obs. Study Perspective

From the above table, the two models we focus on are the first two models, the baseline race/gender/interaction model and the model that handles these variables plus other covariates other than cohort. While Ferrere employs linear analysis over just the rows without missing data, we apply the imputed dataset here to address our questions. This should lead to smaller errors for the coefficient estimates, while also benefitting the individual subset analyses we employ.

We investigate two key subsets of the dataset. The first is the varying "terciles" of mobility. These terciles are ranges of parental education across the respondents. The "low" tercile indicates respondents where parent education is less than twelve, indicating both parents have less than a high school education. The "mid" tercile consists of respondents where at least one of the parents graduated high school, with pared equal to 12 and 13 within this tercile. Finally the "high" tercile consists of parents of respondents where at least one of their parents is college-educated. Each tercile was constructed by looking at the 33rd and 66th quantiles of the parental education variable and forming intervals around these quantiles as bounds.

The second subset of investigation is the effect of race on mobility across various eight-year ranges of birth cohorts. Ferrere talks about how there may be broad differences in mobility by itself across generations. As discussed above, we are curious as to whether any racial differences here can be inferred.

τ_{race} Tercile Estimates

	Model 1		Model 2	
	τ_{race}	SE	τ_{race}	SE
Low	-0.52^{*}	0.207	-0.833^{***}	0.18
Mid	-0.616^{***}	0.156	$-0.282^{\cdot}$	0.154
High	-0.44^{*}	0.19	-0.079	0.167

$^{*}p < .05, ^{**}p < .01, ^{***}p < .001$

τ_{race} Cohort Estimates

	Model 1		Model 2	
	τ_{race}	SE	τ_{race}	SE
1940-1947	-0.576	0.369	-0.716^{*}	0.287
1948-1955	0.674^{*}	0.3	0.02	0.229
1956-1963	0.597^{*}	0.273	-0.077	0.205
1964-1971	0.713^{*}	0.314	0.192	0.237
1972-1979	0.829^{*}	0.382	-0.231	0.287
1980-1987	1.668^{***}	0.482	0.508	0.369
1988-1996	-0.18	0.661	-0.254	0.493

$^{*}p < .05, ^{***}p < .001$

The above table consists of the fitting of Model 1 and Model 2 from our above re-analysis table to the respective subsets of data. Here, to estimate the causal effect of race over mobility, we treat race as a treatment variable and regress

over it to predict mobility as in the model covariates above. If we consider Y to be educational mobility as defined above, Z to be the race variable and implied treatment, X as gender, U as all other non-cohort covariates (pared, paocc10, family16, sibs, incom16), and ε terms the corresponding noise terms of regular OLS, our fitted models over respective intervals of the subsets are:

- Model 1: $Y = Z + X + Z * X + \varepsilon_{XZ}$
- Model 2: $Y = Z + X + Z * X + U + \varepsilon_{UXZ}$

Here, we then employ the outcome regression method as in pages 9-10 of Lecture 6 on observational studies. Given that GSS data consists of observations without an explicit treatment, race is treated as a treatment to infer causal effects from the observational study perspective. Thus, like in outcome regression, we take the estimated causal effect of race on mobility to be the estimate of β_{race} between the two models. Tne analysis over the tercile and birth cohort subsets broadly answer the questions "Has the relationship between educational mobility and race improved across generations?" and "Does this relationship change depending on the education of parents?"

3 Results

The tables tell us that the causal effect of race on educational mobility is largely negative when it comes to past parental education attainment, while it has been somewhat stable over the various decades of the 1900s. The terciles table tells us the causal effect $\hat{\tau}_{race}$ was largely negative across all terciles even with respect to both models. Interestingly, β_{race} is lower in both models within the high tercile of participants. This tells us that race plays slightly less of a role in downplaying educational mobility when at least one of the parents is college-educated.

The cohort results tell us a slight upwards trend of race in affecting educational mobility. It should be noted that the high estimates in the range of 1980-1987 is due to the oversampling of Black respondents in 1982 and 1987 by GSS. Outside of this specific cohort, most of the cohorts in model 1 have a somewhat significant relationship with race, indicating that race plays some consideration in predicting educational mobility when factoring only gender and its interaction for the years of 1948-1979. Model 2 is very different, indicating that race plays little role with little statistical significance when accounting for other covariates, namely those contained in U.

Potential Directions of Future Work

When comparing the work of Ferrare to our analyses, we can note a couple of differences that could provide a base for future work in this project.

One thing to note is that Ferrare made differences between education for a larger subset of individuals, not only examining the difference between Black

respondents and White respondents but also investigating their sex as well. We note here that because we have a defined treatment of just the binary race variable, we largely avoid analysis of sex on mobility as the provided estimation of τ_{race} requires a singular treatment. Adding another "treatment" of sex can overly complicate analysis, but is still a worthwhile covariate that could be explored in depth when considering two-treatment observational study analysis. One method of doing this could be causal designs that employ multiple treatments, especially those used in observational study analysis.

Beyond this relationship, it would be unique to also follow Ferrere's later experiments of investigating gender and race with respect to individual cohorts. Most of the analysis done in his paper focuses on this interaction of gender and sex within the cohorts themselves to infer any generational differences. This would be a unique analysis as well, and I think investigating gender and sex with terciles would also raise interesting questions about past parents education and the relationship between this and gender and race of future generations' educational mobility.

References

1. Crissey, S. R. (2009). Educational attainment in the United States: 2007. US department of Commerce.

2. Davern, Michael; Bautista, Rene; Freese, Jeremy; Morgan, Stephen L.; and Tom W. Smith. General Social Survey 2021 Cross-section. [Machine-readable data file]. Principal Investigator, Michael Davern; Co-Principal Investigators, Rene Bautista, Jeremy Freese, Stephen L. Morgan, and Tom W. Smith. NORC ed. Chicago, 2021. 1 datafile (68,846 cases) and 1 codebook (506 pages).

3. Ferrare, Joseph. Intergenerational Education Mobility Trends by Race and Gender in the United States. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2019-08-27.

4. Krueger, P. M., Tran, M. K., Hummer, R. A., & Chang, V. W. (2015). Mortality attributable to low levels of rducation in the United States. *PLoS ONE, 10(7)*, e0131809.

5. Olshansky, S. J., Antonucci, T., Berkman, L., Binstock, R. H., Boersch-Supan, A., Cacioppo, J. T., . . . Rowe, J. (2012). Differences in life expectancy due to race and educational dif- ferences are widening, and many may not catch up. *Health Affairs*, 31(8), 1803–1813.