Migrant Selection and Socioeconomic Outcomes: Evidence from 19th-Century Sweden

NEKP01: Master Essay II*

Aron Strandberg
Department of Economics
Lund University

October, 2018

Abstract

During the Age of Mass Migration, 30 million Europeans left their home countries and emigrated to the United States. Sweden had one of the highest out-migration rates of the era. Between 1860 to 1920, around 1.3 million people—a quarter of the population—left Sweden to seek opportunity in the United States.

This essay delves into one particular aspect of this historic migration episode. I study the socioeconomic outcomes of Swedish emigrants compared to those who chose to remain in Sweden. Starting with the 1880 Swedish population census, I locate the same individuals 20 years later in either the Swedish or U.S. censuses of 1900. For each year, I use their occupational information to assign each person a standardized socioeconomic status score using the HISCLASS scheme, after which I can compare the outcomes of emigrants and non-emigrants.

In the initial results, I find a significant positive relationship between emigration and socioeconomic attainment, even after controlling for observable pre-emigration characteristics. Then, I use household fixed effects to compare migrants only to their non-migrating siblings. This eliminates the significance of the migration effect, suggesting that the positive socioeconomic outcomes of Swedish emigrants can be explained by their pre-emigration abilities and self-selection.

Supervisor: Kaveh Majlesi

Keywords: Emigration, Socioeconomic status, Self-selection, Sweden

*Presented on Oct 12, 2018. Last revised: Oct 25, 2018.

Contents

1 Introduction

Unlike today, Sweden has not always been a country of immigration. During the Age of Mass Migration (1860-1920), Sweden saw its population leave the country in unprecedented numbers. With around 1.3 million emigrants over the era—a quarter of the population at the time—the Swedish out-migration rate was the third highest in Europe, ranking after only Ireland and Norway (Barton 1994). The vast majority of Swedish emigrants went to the United States which, until the early 1920s, practiced a policy of near-open borders—allowing most Europeans to freely immigrate, work, and eventually naturalize as citizens. Initially, this offer primarily attracted farmers who went to settle in the Upper Midwestern states. In the 1890s, the flow of migrants shifted in favor of urban workers who sought employment in the growing industrial cities (Clemensson 1996).

The micro-level analysis of historical populations is a new field of research that has been made possible by the ongoing digitization of old, often handwritten records such as censuses, passenger lists, and other preserved population schedules. In machine-readable form, these fine-grained data on long-gone populations can now be used by researchers to answer empirical economic questions.

There are several studies that look at different aspects of the historical immigration into the United States. Common topics are the skill distribution of immigrants, the effect of immigration restrictions such as country quotas, and the speed of the immigrant catch-up with native-borns. In this essay, I focus on one particular aspect of migrant outcomes that is comparatively unexplored. The question asked is: How did emigrants fare compared to those who stayed behind in their home country? And, are the observed outcomes primarily a consequence of the migration event itself, or the different characteristics of migrants versus non-migrants?

Using the population censuses of Sweden and the United States, I construct a dataset that links people living in Sweden in 1880 to their location in either one of the two countries in 1900. I then assign each person a socioeconomic status score based on the occupational

information they provide in each of the two years, allowing me to estimate how their relative mobility over 20 years is affected by their status as emigrants or non-emigrants.

Since emigration was not random but largely the result of self-selection, there is reason to believe that the observed success of migrants is the result of some unobservable difference in ability rather than of the migration itself. In the search for causal inference, I construct a sample that groups emigrants with their non-emigrating siblings in Sweden. When controlling for the unobserved heterogeneity that takes place across households, the previously observed positive attainment disappears. This indicates that the observed differences in outcome is a result of selection from particularly capable households.

The rest of the essay proceeds as follows: Section 2 discusses the previous literature and sets up the theoretical framework. Section 3 describes the data sources and the linking procedure. Section 4 describes the empirical strategy. Section 5 presents the results. Section 6 concludes the essay.

2 Background and theory

2.1 Historical background

Although the Swedish emigration era goes back as far as the early 1840s, it would take many years before it took off in large numbers. The first major wave came in 1868-1873 after several years of frost shocks had led to crop failure and famine in Sweden. For most of the 19th century, emigrants mainly consisted of farming families, most of which traveled to settle in the Upper Midwest. A main reason for their attraction to the United States was the promise of cheap arable land, made available through Lincoln's 1862 Homestead Act (Clemensson 1996, Barton 1994).

While the large-scale migration was underway, both Sweden and the United States were in the midst of another concurrent development—the Second Industrial Revolution. This period of rapid industrialization in the final third of the 19th century was socioeconomically

transformative. For workers, it meant a steadily falling share of employment in the agricultural sector and an increased variety of new occupations. Industrialization also strengthened the relationship between education and work, and weakened the generational relationship between the occupations of parents and their children (Treiman 1970). In Sweden, industrialization was spearheaded by the breakthrough of the sawmill industry in the 1870s. Because of this, it was a development that initially took place rurally, and was not as closely linked to urbanization as in many other countries (Karlström 1985).

The introduction of steamships cut both the time and cost of transatlantic travel. When Swedish wages rose by a third between 1865 to 1890, it brought the option of migration within reach of more population strata. As the economic structure of society changed, so did the characteristics of migration. From the 1890s, the previously agricultural migration was replaced by waves of primarily urban workers. This new group of migrants traveled to seek employment in growing industrial cities such as Chicago and New York. Earlier, families had migrated together. The new migrants tended to be younger, unmarried, and traveling alone or with peers (Barton 1994, Pehrson 2014).

Although not with academic rigor, the nature of emigration and emigrant selection was discussed in the contemporary political polemic of the time. One side was of the opinion that those prone to emigration represented the worst of Sweden, and that it even was in the country's favor that they left. One conservative politician named Fredrik Rääf wrote in 1852 that emigration brought Sweden the benefit of *"freeing us from a crowd of rabble, idlers, and restless spirits"*. The newspaper Nya Wermlandstidningen claimed that in general, *"no workers are more lazy, immoral, and indifferent than those who emigrate to other places"*. There were also those that took the opposite view, arguing that those who emigrated represented the best of society. The newspaper Norrlandsposten warned that emigration was driving into exile *"the best part of Sweden's splendid peasantry"* (Barton 1994).

The decades of exodus was made possible only by the policy of open borders that the United States practiced at the time, allowing most Europeans to freely immigrate into the

country with little restriction[1]. In the early 20th century— possibly explained by the nativist reaction to southern and eastern European immigration—the free-movement zeitgeist came to an end. Congress limited the inflow of immigrants by enacting a series of increasingly strict quota laws in the 1920s, after which the Age of Mass Migration was over (Abramitzky & Boustan 2017).

2.2 Literature and theoretical approach

The micro-level analysis of historical populations is a new area of economic research that has opened up only in recent years following the increasing amount historical records—often handwritten censuses with millions of observations— being digitized and made available for research.

Projects such as HISCO (Van Leeuwen et al. 2002) and its derivatives are dedicated to creating standardized measures of occupations, status, and stratification using the information provided in historical records. The North Atlantic Population Project (Minnesota Population Center 2017) facilitates research by collecting and harmonizing historical census data into a unified format that eases comparability across different censuses. There are also many other non-census records that have been digitized primarily for genealogical purposes, but have proven to be just as useful in academic research.

Although many—the author of this essay included—would consider the musings on history interesting for their own sake, the field also plays an important role in economics, as it allows researchers to study empirical economic decisions and incentives under institutional settings that are no longer available in the present day. In the migration context, it makes it possible to study migration-related outcomes under an era of low barriers and comparatively undistorted migration incentives.

The pioneering works by Wegge (1998, 2002) are the first that study 19th century migrant selection using individual-level microdata. Wegge analyses a sample of German emigrants

[1]The rules were less permissive for other groups. For example, the exclusion acts of 1882 and 1908 specifically prohibited Chinese and Japanese immigration.

from Hesse-Cassel in the 1850s and finds that migrants to the United States were disproportionally drawn from the middle of the occupational skill distribution.

Abramitzky et al. (2012) construct a dataset consisting of Norwegian migrants and non-migrants in 1900 that can be linked back to their childhood homes in 1865. After assigning each person the mean earnings of their occupational category, they estimate a positive return to migration. The authors then use household fixed effects to compare migrants only to their brothers who stayed behind in Norway. They find that migrants from urban areas were negatively selected on skill, whereas rural migration was more ambiguous.

Detailed earnings data are rarely available for this era. In lieu of this, social stratification researchers have developed alternative measures to infer information about socioeconomic trends. These, unlike earnings, do not tell about absolute monetary returns, but can still capture important information about socioeconomic and occupation-based trends.

One commonly used measure of socioeconomic status is HISCLASS (Van Leeuwen & Maas 2011). It is an occupational class scheme that sorts occupations into twelve broad socioeconomic classes based on a set of criteria such as skill level, the degree of supervision, if the work is manual or non-manual, and economic sector. HISCLASS has been created with the purpose of allowing consistent comparison across time and place in historical records. Table 1 provides some examples of occupations and their corresponding HISCLASS score.

Dribe et al. (2017) use HISCLASS to study socioeconomic heterogamy of women in Sweden. They link the censuses of 1880, 1890 and 1900, and follow women from their parental homes to their marital households. They find a strong relationship between internal migration within Sweden and hypergamy, highlighting the importance of migration for social mobility during the industrialization era.

Massey (2016) estimates the effect of the U.S. Quota Act of 1921 on skill by comparing the quota-affected groups with unaffected Canadian immigration. She finds that the quota resulted in migrants with higher skill. Similarly, Wulfers (2018) uses HISCLASS to study how the skill composition of U.S. immigrants was affected by the quotas of the 1920s. Using

Table 1: HISCLASS occupations

Class	HISCLASS	Examples
1	Higher managers	Banker, hotel owner
2	Higher professionals	Laywer, architect
3	Lower managers	Postmaster, sea captain
4	Lower professional, clerical, and sales personel	Merchant, telegraph operator
5	Lower clerical and sales personel	Store clerk, salesman
6	Foremen	Housekeeper, butler
7	Medium-skilled workers	Carpenter, baker
8	Farmers and fishermen	Farmer, cattle raiser
9	Lower-skilled workers	Coal miner, house painter
10	Lower-skilled farmworkers	Lumberman, hunter
11	Unskilled workers	Laborer, domestic servant
12	Unskilled farmworkers	Farm laborer, gardener

passenger lists from ships leaving Bremen, he finds that the increasingly strict quotas led to an increase in the skill level of immigrants, first as a shift from agricultural workers to manufacturing workers, and then to professional workers.

It is possible that people who decided to emigrate were not representative of their general demographic groups, but were self-selected on some more intrinsic quality. Knudsen (2017) constructs proxy variables for individualism based on the uniqueness of names and the relative use of singular pronouns, and finds emigrants displayed a higher degree of individualism. Due to the sheer size of the migration flow, this had consequences not only for emigrants themselves, but also for the place from which they left. Knudsen shows that areas with more emigration turned more collectivist as a result.

In this essay I focus on the outcomes of emigrants themselves, relative to their initial starting position that is observed 20 years prior. One one hand, one can imagine that migration constitutes a treatment that sets individuals on a new trajectory of socioeconomic attainment, perhaps due to the better opportunities that the new country has to offer. The other hypothesis—more in line with the idea of Knudsen—is that the people that have a stronger propensity to emigrate also possess an exceptionalism that influences their outcomes. In the counterfactual scenario where migration is not an option, it is possible that the

same individuals turn out just as successful. I return to the method used to analyse these perspectives in section 4.

3 Data

3.1 Swedish census data

The decennial Swedish censuses between 1880 to 1910 have been digitized by the Swedish National Archives and are made available in the NAPP database (Minnesota Population Center 2017). The 1880 census counts 4,624,825 people in 2,587 parishes, and the 1900 census counts 5,200,111 people in 2,639 parishes.

For each person living in Sweden at the time, the censuses list information such as name, age, sex, birth year, birthplace, and occupation. Within each household, NAPP have added codes that indicate the relationships between the members. I use these relationship codes to add information about the father's occupation and to create the groups of siblings that are analyzed in the fixed-effects analysis.

Geographical information such as parishes have been harmonized for compatibility across censuses. Occupations have been standardized with HISCO occupational codes, which allows easy mapping to HISCLASS using the official conversion table by Van Leeuwen & Maas (2011)

A peculiarity of the Swedish population records is that children lack surnames of their own. The task of constructing their probable adult surname is complicated by the fact that Sweden had two concurrent surname practices in use at the time. People would either take their father's last name as a family name, or take a patronymic based on their father's first name. The son of a Lars Svensson could therefore get the surname Larsson or Svensson, and a daughter could also possibly get the surname Larsdotter. After parsing out the -sson and -dotter suffixes, the linking procedure is therefore run twice to search for each of these alternatives.

Swedish censuses were carried out very differently from most other countries. Instead having enumerators gather information through interviews, the vicars or parish priests were tasked with keeping track of the population in parish books, from which the information was provided to the census enumerators when requested (Scalone & Dribe 2010). As a result, the information stays relatively consistent between years, with the main variation being in the inconsistent spelling of names. In many census linking studies it is necessary to allow age discrepancies of up to five years to be able to locate individuals across different records. In Sweden, such discrepancies are largely non-existent.

3.2 Emigration data

The Emibas (2005) database contains the records of 1.1 million people from more than 2,300 parishes who emigrated from Sweden between 1830 to 1940. Around 80 percent of all Swedish emigrants are listed in these records.

The data comes from the local parish books, which took note of the date when people departed. In addition to the same information that is available in the Swedish censuses, the emigration records just like the U.S. census also provides the month of birth.

According to these records, 582,566 people emigrated during the period of interest between 1880 to 1900. 86 percent of these state the United States or North America (but not Canada) as their destination. Of the rest, many emigrated to nearby countries such as Denmark and Germany, but many destinations are also illegible.

3.3 U.S. census data

A subset of the 1900 U.S. Federal Census was obtained from Ancestry.com. This contains all individuals in the census that have listed their country of birth as Sweden and year of immigration between 1880 to 1900. In total, it counts to 408,771 people.

Each working person in the records have provided their occupation in the form of a text string. Unlike the NAPP-provided records, these do not have HISCO occupation codes.

Although there are more than 33,000 different occupations provides, most are just slight spelling variations. To standardize them, I retrieve the 1880 U.S. census from NAPP which has been fully coded to HISCO. I then fuzzy-match these occupations to my own sample, allowing a Jaro-Winkler distance of up to 0.15.

3.4 Linking procedure

The goal is to for each individual in the 1880 Swedish census, try to find the same individual again in one of the 1900 censuses.

Unlike the Swedish datasets which have parishes of birth, the origin in the U.S. census is not more specific than just "*Sweden*". When Swedes arrived in the United States, they would also normally anglicize their names. For example, someone named Johannes Svensson would become John Swanson, and Maria Pehrsdotter would become Mary Pearson.

To overcome these hurdles, I start by matching individuals to the emigration data. Then in the second step, I link the confirmed migrants to the U.S. census with a very exact search on their arrival year, and month and year of birth. The use of this two-step process instead of linking directly to the U.S. census brings down the search for the correct person to just a few potential matches. The median parish had only 1,232 residents in 1880. Compared to e.g. Abramitzky et al. (2012) which searches for unique matches nation-wide, this approach vastly brings down the number of potential matches and might increase the accuracy.

Abramitzky et al. (2018) have created a fully automated method for linking historical datasets in a way that minimizes both false positives and false negatives, using a expectation–maximization (EM) algorithm. I largely follow the procedure outlined the data appendix of this paper, with a few changes to fit my data.

With the Swedish censuses and emigration database combined, the linking is done as follows. First, I parse out -sson and -sdotter from the endings of last names. When the name of person's father is available, one patronymic and one family surname is constructed.

Then, I divide the data into smaller datasets using the blocking variables parish of birth,

sex, and the first letters of the first name and last names. Potential matches are compared only within each block. This is done to keep the distance matrices within reasonable computational limits. Similar letters are merged to allow for spelling discrepancies, such as C and K, or W and V.

Within each block, the Jaro-Winkler distance between first names and last names are calculated respectively. The distance between each person's birth year is calculated and deemed a potential match if the difference is at most one year.

These three distances are then combined into a combined measurement of similarity. The weight of each component is automatically calculated using an expectations-maximization algorithm provided with Abramitzky et al. (2018). The procedure is run twice. Once using constructed patronymic names, and once with constructed family names.

To confirm a match, we require that the best potential match is above a certain similarity threshold, and that the second-best match is below a threshold. An arbitrary choice has to be made on which thresholds to use. In the data appendix of Abramitzky et al. (2018), different sets of parameters are recommended that correspond to lenient and conservative samples. I use the lenient parameters for the sample presented in this essay.

The same procedure is then run again to link the emigrant dataset to the U.S. census. The only difference is that the blocking variables are now the first letters of the names, sex, year of immigration, and month of birth.[2]

As a result, I am able to link 1,706,775 non-migrants from the first Swedish census to the second one. These will constitute the control group. Of the 95,166 people in the U.S. census that are also found in the Swedish emigration records, 48,423 can be linked to the first Swedish census. These migrants constitute the treatment group. Combined, the total sample consists of 1,755,198 individuals. Summary statistics are found in table 2.

In the censuses, women rarely have an occupation of their own. For married women, it

[2]The restrictiveness of these blocking variables makes the good matches unique already from the start. The sample is very similar to if a simple merge is done on blocking variables and non-unique matches sorted out.

Table 2: Summary statistics

Statistic	N	Mean	St. Dev.	Min	Max
Age	1,755,198	44.403	17.168	20	116
Sex	1,755,198	0.495	0.500	0	1
Emigrant	1,755,198	0.028	0.164	0	1
HISCLASS 1900	855,845	8.826	2.485	1	12
Father's HISCLASS 1880	1,555,854	8.645	2.296	1	12
Parish size	1,755,198	3,891.688	4,717.266	30	34,298
Year of immigration	48,423	1,887.771	4.903	1880	1900
Urban	1,755,198	0.127	0.333	0	1
Lives on farm	1,755,198	0.474	0.499	0	1
Number of siblings	1,755,198	1.444	1.972	0	16
Birth order	828,280	2.508	1.590	1	15
Is oldest sibling	828,280	0.342	0.474	0	1
County	1,755,198	12.609	6.506	1	25

is virtually non-existent. In the literature, such as in the earlier mentioned study by Dribe et al. (2017), a common way to study the outcomes of married women is to use the status of their spouse as a proxy for their own socioeconomic status. The same regressions can then be used to indirectly study their hypergamy.

The linking of women is made complicated by the fact that women in Sweden generally did not change their surnames upon marrying whereas in the United States they did. There will also be matching skewness in favor of those who married before departing, which can influence the results. This deserves a study in its own right, and regressions with women's outcomes are therefore omitted from the main results here. The results from the subset of married women is included in the appendix. These show similar estimates as for men, but should be considered tentative and potentially more bias-prone.

4 Empirical specification

The goal of the analysis is to estimate the relationship between migration and socioeconomic status. The starting point is the following equation.

$$SES1900_i = \alpha + \beta_1 Emigration_i + \beta_2 SES1880_i + \mathbf{X_i'}\beta_{\mathbf{X}} + \varepsilon_i \tag{1}$$

The outcome variable $SES1900_i$ is defined as the person's *HISCLASS* score in 1900 multiplied by (-1). The negative term makes the sign in the results more intuitive, since HISCLASS itself decreases with higher status[3].

$Migration_i$ is a dummy variable that indicates if person i is living in the United States in 1900. Its estimated coefficient is the main point of interest in this essay. In some of the regressions presented in the results, the number of years spent in the United States—between 0-20—is used as an alternative variable.

$SES1880_i$ is the pre-treatment socioeconomic status of person i in 1880. It is constructed like the dependent variable. This is the most important control variable, as it is the main predictor of a person's future status. Children and many young adults do not have occupations of their own to derive their socioeconomic status from. In the regressions that look at them separately, the father's *HISCLASS* is used as a proxy variable for their status.

Finally, $\mathbf{X_i'}$ is a vector of pre-treatment control variables. It includes likely influencing factors such as age, family characteristics, and urban versus rural status. It also includes dummy variables indicating the county of birth.

4.1 Identification strategy

To make things complicated for later migration researchers, emigrants were not randomly drawn from the population, but made the choice to leave via self-selection. The main threat

[3]Although HISCLASS is intrinsically an ordinal variable rather than a nominal one, the validity of its used as an approximate nominal measure in linear regression has been established in earlier literature. See for example Wulfers (2018). A more rigorous alternative for HISCLASS is ordinal logistic regression, but I stick to the former in this essay due to the more easily interpreted results.

to a causal interpretation is the existence of some underlying unobserved ability that influences both the propensity to migrate and the chance of socioeconomic success (or failure). Following Abramitzky et al. (2012), I create a sample that consists only of matched brothers that share the initial characteristics from their 1880 household but then potentially differ in 1900 outcomes. Only 1880 households that have at least two matched brothers with non-missing socioeconomic information are kept in this model.

For the household fixed effects model, the following equation is considered.

$$SES1900_i = \alpha + \beta_1 Emigration_i + \mathbf{X_i'}\beta_{\mathbf{X}} + \alpha_j + \nu_{ij} \tag{2}$$

In this equation, the individual error term is decomposed into two components. Here, α_j is the error component that is shared between siblings that belong to the same household j and ν_{ij} is the error component that is unique to the individual. When household fixed effects are included in this model, the bias that arises due to differences in α_j across households is eliminated. The vector $\mathbf{X_i'}$ includes only variables that vary within households, such as age and sibling order.

5 Results

5.1 Initial results

Four different estimates of equation 1 are presented in table 3 below. The sample in this table consists of all men that, in both years, have an occupation of their own to derive their socioeconomic status from. Columns (1) and (3) are baseline estimations using different treatment variables. Columns (2) and (4) are expanded versions of the respective previous column. All regressions include county fixed effects and use robust standard errors.

In the first two colums, we see a positive effect of emigration on the socioeconomic outcome. For example, the point estimate in column (1) shows that migration on average

is associated with a 0.717 points higher status in 1900. The point estimate in column (2) is somewhat lower.

The initial status is positively correlated with the outcome status, as would be expected.

Both *Age* and *Age²* are included to allow a quadratic relationship, which is commonplace in this context. Contrary to expections, the results here show a U-shaped relationship between age and the outcome variable.

Parish size is defined as the square root of the population size in 1880. The variable is closely related to the *Urban* dummy variable which indicates if the parish is defined as rural or part of a city. Both estimates are positive in their relationship with later socioeconomic status.

In columns (3) and (4), I look at an alternate specification, where the emigration dummy is replaced by a variable *Years in US* that indicates how many years the person have lived in the United States, if any. The variable is also included in squared form. This is done on the assumption that emigration could carry a status penalty in the short run, before the person has been able to catch up in the new country. The number of years in the United States are found to be positively correlated with the socioeconomic outcome, but the negative short-term effect is non-significant[4].

In table 4, the sample is restricted to men aged between 20 to 50 years old in 1900. Unlike men outside this age range, they were likely to live with their parents and siblings in 1880, which is required to be able to link them together. The father's status in 1880 is used as a proxy variable for initial status in these regressions.

All the emigration-related point estimates are smaller in the results of table 4, as is the proxy for initial status. A comparison of R^2 between tables 3 and 4 shows that using the person's own SES in 1880 rather than the father's captures much more of the variation, as is expected.

The age estimates do here have their intuitive signs, showing a positive but decreasing

[4]I also try using the *year of immigration* variable to restrict the sample to people who were adults when arriving in the United States. The results are very similar and not presented here.

effect of age. The discrepancy from the table 3 could be the result of the difference in age distribution between the two, where this sample is much younger.

Table 3: Results

	Dependent variable:			
	SES 1900			
	(1)	(2)	(3)	(4)
Emigrant	0.717***	0.495***		
	(0.027)	(0.027)		
Years in US			0.049***	−0.002
			(0.002)	(0.010)
Years in US^2				0.002***
				(0.001)
SES 1880	0.611***	0.633***	0.611***	0.626***
	(0.001)	(0.002)	(0.001)	(0.002)
Age		−0.111***		−0.108***
		(0.003)		(0.003)
Age^2		0.001***		0.001***
		(0.00002)		(0.00002)
Parish size		0.003***		
		(0.0001)		
Is urban				0.374***
				(0.010)
Constant	−2.840	0.776	−2.839	0.689
	(0.018)	(0.082)	(0.018)	(0.082)
County FE	Yes	Yes	Yes	Yes
Observations	359,560	359,560	359,560	359,560
R^2	0.404	0.418	0.404	0.420
Adjusted R^2	0.404	0.418	0.404	0.420

Note: *p<0.05; **p<0.01; ***p<0.001
All regressions use HC2 robust standard errors

5.2 Household fixed effects

In the final part of the results, I turn to the household fixed-effects estimation of table 5.

Just like in the previous table that was presented for comparability, the sample is restricted to men aged between 20-50. In order to contribute to the within-household estima-

Table 4: Results: Men aged 20-50

	Dependent variable			
	SES 1900			
	(1s)	(2)	(3)	(4)
Emigrant	0.210***	0.212***		
	(0.017)	(0.017)		
Years in US			0.025***	0.024***
			(0.001)	(0.001)
SES 1880 (father)	0.407***	0.385***	0.407***	0.372***
	(0.002)	(0.002)	(0.002)	(0.002)
Age		0.085***		0.085***
		(0.004)		(0.004)
Age^2		−0.001***		−0.001***
		(0.0001)		(0.0001)
Parish size 1880		0.009***		
		(0.0002)		
Urban 1880				1.021***
				(0.012)
Number of sib.		0.050***		0.058***
		(0.002)		(0.002)
Constant	−5.104	−8.102	−5.105	−7.898
	(0.024)	(0.071)	(0.024)	(0.069)
County FE	Yes	Yes	Yes	Yes
Observations	388,752	388,752	388,752	388,752
R^2	0.146	0.166	0.147	0.175
Adjusted R^2	0.146	0.166	0.147	0.175

Note: *p<0.05; **p<0.01; ***p<0.001
All regressions use HC2 robust standard errors

tion, a 1880 household must contain at least two brothers who are matched to 1900. All observations that fail this criteria are dropped.

In columns (1) and (2), the sample is restricted to the matched-brother data, but run without the household fixed effects. The estimates look similar to the ones found in the previous table.

Variables are included that indicate the number of siblings and whether the person is the oldest male sibling. Based on reasoning in Abramitzky et al. (2012), the expectation and reason for including them was that having fewer siblings, or being the oldest son in particular, increases the likelihood of inheritance of the parents' property (and in some cases, their occupations), something that can have an effect on both social mobility and the likelihood to migrate. The signs of these estimates are contrary to these expectations, however. Having more siblings is associated with a worse socioeconomic outcome, and being the oldest brother is associated with a more positive one.

In columns (3) and (4), household fixed effects are added. These absorb the part of the error term that is fixed within the household. That is, all bias that arises due to aspects in family background that potentially correlate with the probability to migrate and with socioeconomic outcomes. There are 5,451 households with at least one migrant and one non-migrant brother who contribute to the estimation of the treatment effect in the model. Still, siblings can differed in their unobserved personal abilities within a family. This part of the error term remains uncontrolled for here. With household fixed effects included in columns (3) and (4), the significance of the migration estimates disappears. Apart from the fixed effects themselves, age remains as the sole variable with explanatory power.

Table 5: Results: Sibling fixed effects

	(1)	(2)	(3)	(4)
Emigration	0.172***	0.143***	0.0719	0.0360
	(0.0293)	(0.0290)	(0.0434)	(0.0432)
SES 1880 (father)	0.434***	0.424***		
	(0.00324)	(0.00327)		
Age		0.159***		0.145***
		(0.00730)		(0.00930)
Age^2		-0.00176***		-0.00154***
		(0.000110)		(0.000144)
Number of sib.		-0.0161***		
		(0.00378)		
Is oldest sib.		0.0308*		0.000969
		(0.0157)		(0.0199)
Constant	-5.058***	-8.226***	-8.820***	-11.76***
	(0.0301)	(0.121)	(0.00561)	(0.149)
Observations	144943	144943	144943	144943
Groups			90126	90126
Migrant-stayer groups			5451	5451
Sibling FE	No	No	Yes	Yes
Adjusted R^2	0.148	0.163	0.420	0.429

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

6 Concluding remarks

The purpose of this study has been to explore the effect of migration on later socioeconomic outcomes by studying the outcomes of Swedish emigrants to the United States at the height of the Age of Mass Migration.

I initially find that emigrants on average saw larger increases in their socioeconomic status, as measured by their and their father's HISCLASS scores, even when controlling for a number of observable characteristics.

To better underpin causality, I then estimate a model with sibling fixed effects, controlling for all variation that takes place across households. In these results, the previously found significance of migration disappears. This could be an indication that the cause of relative success of emigrants lies in some unobserved ability that is shared within the household, rather than in the migration event itself.

The findings of this essay are meant to shed some light on the nature of Swedish emigration and its socioeconomic consequences, adding to the literature of literature of studies using micro-level historical data to study past economic outcomes.

Much remains to be tackled in future research. The ongoing digitization of more censuses will allow researchers to study individual outcomes over longer stretches of time. With clever techniques, more information can be revealed from the records that are already available, for example, by using the uniqueness of names or heaping of birthdays around round numbers to infer information about individuality and numeracy, respectively. This essay and the previous literature have only scratched the surface of a topic where there are many avenues for future exploration.

References

Abramitzky, R. & Boustan, L. (2017), 'Immigration in american economic history', *Journal of economic literature* **55**(4), 1311–45.

Abramitzky, R., Boustan, L. P. & Eriksson, K. (2012), 'Europe's tired, poor, huddled masses: Self-selection and economic outcomes in the age of mass migration', *American Economic Review* **102**(5), 1832–56.
URL: *http://www.aeaweb.org/articles?id=10.1257/aer.102.5.1832*

Abramitzky, R., Mill, R. & Pérez, S. (2018), Linking individuals across historical sources: a fully automated approach, Technical report, National Bureau of Economic Research.

Barton, H. (1994), *A Folk Divided: Homeland Swedes and Swedish Americans, 1840-1940*, Studia multiethnica Upsaliensia, Southern Illinois University Press.
URL: *https://books.google.se/books?id=9m1zOOpFTwwC*

Clemensson, P. (1996), 'Göteborgska källor till den stora utvandringen', *Dokumentet* **3**.

Dribe, M., Eriksson, B. & Scalone, F. (2017), 'Migration, marriage, and social mobility: Women in sweden 1880-1900', *Lund Papers in Economic Demography* **2017**(1).

Emibas (2005), *Emigrantregister för Sverige: Version 1.00 [CD]*, Svenska emigrantinstitutet, Växjö, Sweden.

Karlström, U. (1985), *Economic growth and migration during the industrialization of Sweden. A general equilibrium approach*, Handelshoegskolan.

Knudsen, A. S. B. (2017), 'Historical individualism: Selective migration and cultural persistence'.

Massey, C. G. (2016), 'Immigration quotas and immigrant selection', *Explorations in Economic History* **60**, 21–40.

Minnesota Population Center (2017), *North Atlantic Population Project: Complete Count Microdata: Version 2.3 [dataset]*, University of Minnesota, Minneapolis, MN.
URL: *https://doi.org/10.18128/D040.V2.3*

Pehrson, L. (2014), *Den nya staden: Utvandringen till Amerika II*, Albert Bonniers Förlag.
URL: *https://books.google.se/books?id=6cGKAwAAQBAJ*

Scalone, F. & Dribe, M. (2010), 'Socioeconomic status and net fertility in the demographic transition: Sweden in 1900–a preliminary analysis', *Popolazione e storia* **11**(2), 111–132.

Treiman, D. J. (1970), 'Industrialization and social stratification', *Sociological inquiry* **40**(2), 207–234.

Van Leeuwen, M. H. & Maas, I. (2011), *HISCLASS: A historical international social class scheme*, Universitaire Pers Leuven.

Van Leeuwen, M. H., Maas, I. & Miles, A. (2002), *HISCO: Historical international standard classification of occupations*, Leuven Univ Pr.

Wegge, S. A. (1998), 'Chain migration and information networks: Evidence from nineteenth-century hesse-cassel', *The Journal of Economic History* **58**(4), 957–986.

Wegge, S. A. (2002), 'Occupational self-selection of european emigrants: Evidence from nineteenth-century hesse-cassel', *European Review of Economic History* **6**(3), 365–394.

Wulfers, A. A. (2018), Skill selection and american immigration policy in the interwar period, Technical report, University of Oxford, Department of Economics.

Appendix

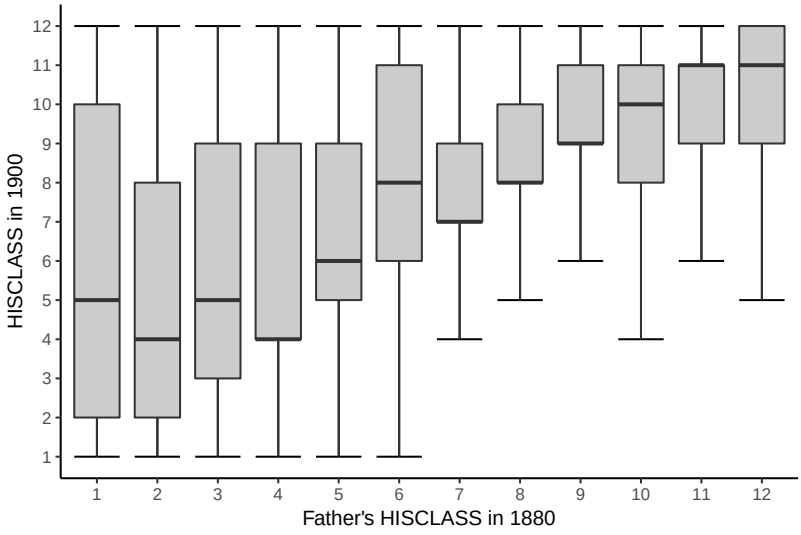

Figure 1: Men aged 20-50: Outcome distribution, by father's SES

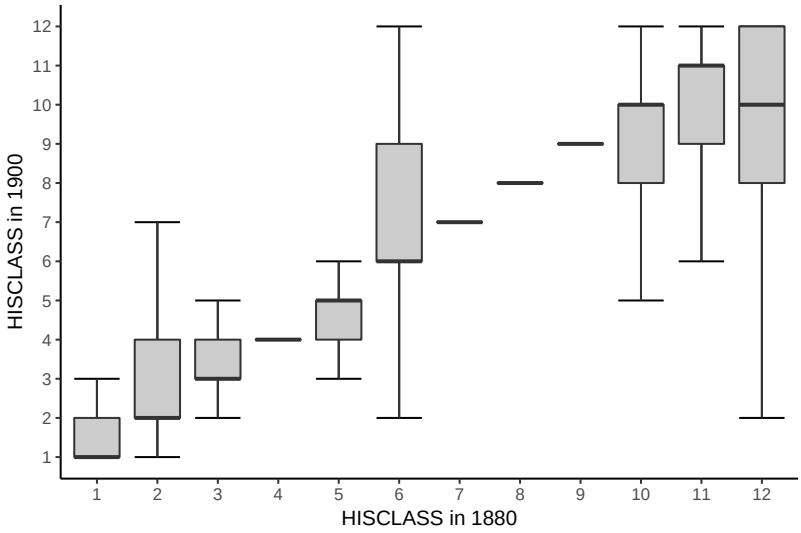

Figure 2: All men: Outcome distribution, by own SES

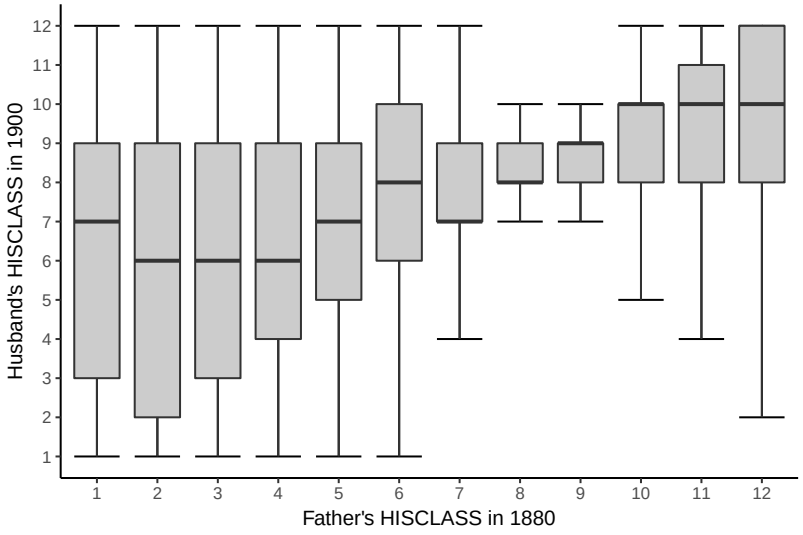

Figure 3: Women aged 20-50: Outcome distribution, by father's SES

Table 6: Results: Married women aged 20-50

	Dependent variable:			
	SES 1900 (spouse)			
	(1)	(2)	(3)	(4)
Emigrant	0.466***	0.434***		
	(0.026)	(0.026)		
Years in US			0.035***	0.033***
			(0.002)	(0.002)
SES 1880 (father)	0.308***	0.297***	0.308***	0.284***
	(0.002)	(0.002)	(0.002)	(0.002)
Age		0.015**		0.015**
		(0.005)		(0.005)
Age Squared		0.00000		−0.00000
		(0.0001)		(0.0001)
Parish size 1880		0.008***		
		(0.0002)		
Urban 1880				0.983***
				(0.017)
Number of sib.		0.088***		0.094***
		(0.002)		(0.002)
Constant	−6.234	−7.463	−6.235	−7.317
	(0.033)	(0.099)	(0.033)	(0.098)
County FE	Yes	Yes	Yes	Yes
Observations	262,786	262,786	262,786	262,786
R^2	0.103	0.114	0.103	0.123
Adjusted R^2	0.103	0.114	0.103	0.123

Note: *p<0.05; **p<0.01; ***p<0.001
All regressions use HC2 robust standard errors