Europe's tired, poor, huddled masses: Self-selection and economic outcomes in the age of mass migration

Ran Abramitzky, Leah Platt Boustan, Katherine Eriksson

Web Appendix

I. Matching between the 1865 and 1900 Censuses

Our goal is to match Norwegian-born men in 1900 to their childhood households in the 1865 Norwegian Census. We use two sources in 1900: the Norwegian Census, which is archived by the North Atlantic Population Project (NAPP), and a complete roster of Norwegian immigrants living in the US, which we compiled from the genealogy website Ancestry.com. Because over 95 percent of emigrants from Norway settled in the United States, these two sources contain nearly all Norwegian-born men who survived to 1900 (Ferenczi and Willcox, 1929).

Our baseline method ("**Match 1**") uses an iterative matching strategy pioneered by Ferrie (1996).

We describe this procedure in detail:

(1) We identify 257,767 Norwegian men between the ages of 3 and 15 in 1865. 71,644 of these men are unique by first name, last name and birth year in 1865.

(2) We standardize all first and last names in both datasets to address orthographic differences between phonetically equivalent names using the NYSIIS algorithm (see Atack and Bateman, 1992).

(3) We match unique observations in 1865 forward to 1900 using an iterative procedure. We start by looking for a match by name and exact birth year. If we find a *unique* match here, we stop and consider the observation "matched." If we find multiple matches for the same birth year, the observation is thrown out. If we do not find a match at this

first step, we try matching first within a one-year band (older and younger) and then with a two-year band around the reported birth year. If neither of these attempts produces a match, the observation is considered to be "unmatched."[1]

This procedure generates a sample of 2,613 migrants and 17,833 non-migrants. We achieve a forward match rate of 29 percent, which is comparable to Long and Ferrie's (forthcoming) forward match rate of 22 percent within the United States over a similar 30 year period (1850-80).[2]

II. Causes of match failure

We match 30 percent of the 71,644 potential matches from 1865 to 1900, fail to match 4 percent due to name-age combinations that are not unique in 1900 and fail to match the remaining 66 percent due to name-age combinations that cannot be found in 1900. The number of successful matches and match failures by cause are reported in Online Appendix Table 1. Failures due to name-age doubles in 1900 likely reflect age misreporting. Indeed, when we expand our age band to 10 years, we locate an additional 4,328 individuals, suggesting that 6 percent of potential matches misreported their age by a few years (=4,328/71,644). All of our results are qualitatively similar if we expand our age band to 10 years. Match failures due to missing name-age combinations could be due either to mortality between 1865 and 1900 or to name changes or transcription error. We use US mortality rates by age and sex in 1900 to calculate an expected 35-year mortality

[1] We restrict our attention to men who are at least three years old in 1865 to ensure that all observations can match to a two-year age band around the reported age.

[2] If we instead conducted a backwards match from 1900 to 1865, we achieve a match rate of 23 percent, a weighted average of 11 percent for men living in the US in 1900 and 25 percent for men living in Norway. The differential match rate between migrants and non-migrants could be due to the practice of anglicizing one's name upon arrival in the US, a factor that we consider in the next section.

rate of 25 percent. We conclude that mortality can account for around 40 percent of the match failures due to missing name-age combinations.

Online Appendix Table 2 compares the characteristics of our matched sample to subgroups of the unmatched population in 1865. There are two patterns worth noting. First, men who are unique by name and age in 1865 are substantially more likely than others to live in urban areas and in a household with a higher socio-economic status, regardless of whether they eventually can be matched to 1900. This comparison is consistent with our claim that having an uncommon name is associated with being from a higher-status (more cosmopolitan) background. Second, even among the subset of men who are unique in 1865, men who match to 1900 are more likely to live in a high-status household. This fact is likely due to the fact that many of the causes of match failure, including mortality and age misreporting, are negatively associated with socio-economic status.

III. Sources of occupation-income data

Men in the United States are assigned income data based on their occupation from the US 1901 Cost of Living Survey. Men in Norway are matched to mean income-by-occupation tabulations for the year 1900 published by Statistics Norway and other sources (Haines and Preston, 1991; *Statistik Aarbog*, 1900; Grytten, 2007).[3] The 1901 Cost of Living Survey reports income information for more than 300 occupations in the US.[4] Our dataset contains individuals representing one hundred and eighty-nine occupational categories. We convert

[3] *Statistics Norway* reports daily wage rates. We convert these wage rates into annual earnings figures by assuming that Norwegians worked six-day work-weeks and were unemployed for 0.66 months during the year (= 297 days of work per year, on average). Our estimate for months spent unemployed is based on reported unemployment for Norwegian migrants in the 1900 US Census.

[4] For men living in the US, we code occupation by hand using the digital images of Census manuscripts available on Ancestry.com.

Norwegian wages to real, PPP-adjusted US dollars using the 1900 exchange rate and price levels reported in Grytten (2004).

The 1901 Cost of Living survey may overstate the return to migration both because the survey was conducted in urban areas and because the majority of survey respondents were native born. We address these concerns by considering alternative sources of earnings data in the US. First, we calculate earnings by occupation from the 1915 Iowa Census, which better represents the urban/rural divide in our US sample.

IV. Estimating income for farmers, fisherman and white collar workers

Standard sources do not report information on earnings for owner-occupier farmers in either the United States or Norway. We follow Mitchell, et al. (1922) in estimating the net earnings of owner-operator farmers from farm revenue and expenditures data. For the United States, we use data on farmers in Minnesota, the most common state of residence in our sample, from the 1900 Census of Agriculture. For Norway, we use data for the total value of farm products for the 1900 harvest found in the 1907 Census of Agriculture (*Jordbruksteljinga*).

The 1907 Census of Agriculture reports the total value of farm product, rather than average value per farm. According to the 1900 Census, total farm output in Norway is produced by 133,400 owner-operators, 73,200 farm laborers, 24,500 tenant farmers and 35,800 individuals who report being "farmers and fisherman." To estimate the earnings of owner-occupiers, we need to subtract the value added by tenant farmers and the composite "farmer and fisherman" category; farm labor is already accounted for on the expenditures side of the ledger. The average farm laborer earned $185 a year (US $1900). We assume that, with free mobility, tenant farmers would have earned the same amount as farm laborers (in

expectation). Therefore, we subtract $4.5 million (=24,500 · $185) from the total value of farm product. Furthermore, we assume that men who report being "farmers and fisherman" earn a subsistence living and eat what they produce. Thus, we divide total farm product less $4.5 million by the number of owner-operators.

The 1906 Statistics Annual (*Statistik Aarbog*) reports the total value of cod, herring, mackerel, salmon, merlan, lobster and oysters sold in 1900. The 1910 volume *Gages Annuels des Domestiques et Salaires des Ouvriers* indicates that, in deep-sea fishing expeditions, fishermen typically received 35-55 of the catch. We divide this total by the 41,680 fisherman in the 1900 Census.

With the exception of primary school teachers, we could not locate income data for white collar workers in Norway in 1900 (24 percent of the labor force). We assign these workers the relevant income level from the United States deflated by the average Norway-US income gap. If the return to skill was higher in the United States than in Norway, this procedure will understate the total return to migration.

V. Comparing matched sample to full population

Online Appendix Table 6 compares the attributes of men in the primary matched sample to the Norwegian population in the same age range in the 1865 Census, while Online Appendix Table 7 compares matched migrants to Norwegian-born men living in the United States and matched stayers to men living in Norway in 1900.[5]

[5] For these tables and the remainder of the analyses in the paper, we drop men who lived in group quarters in 1865 (1,676 men in our matched sample) because we are unable to reconstruct aspects of their childhood households.

By construction, men with uncommon names are more likely to be successfully linked between Censuses. Table A6 shows that the median rural man in the population shared his first and last name with 121 others, while the median urban man shared his name with 11 others. Unsurprisingly, name frequency in the matched sample is substantially lower than that of the population, with rural men sharing their name with 7 others and urban men with only 2 others in the Norwegian population.

According to Table A6, only 14 percent of Norwegian men lived in an urban area in 1865 compared with 26 percent of our matched sample.[6] Within urban areas, matched men are 10 percentage points more likely to live in a household whose head holds an occupation with above-median earnings and twice as likely to live in a household with some assets, defined as owning a business or serving as a master craftsman in an artisanal workshop. In rural areas, matched men are also drawn from higher socio-economic status households although this difference is not as pronounced.: men in our matched samples earn around 4 percent more than the comparable population in 1900, both among migrants to the US and men who remain in Norway (Table A7). As we expected, having an uncommon name appears to be a proxy for urban location and socio-economic status, perhaps because urban households used a wider array of given names (Gjerde, 1985, p. 48).

[6] Norwegian households were defined by the Census as urban if their municipality of residence was considered to be a town. However, many towns contained agricultural land on their periphery. Therefore, the urban designation likely includes some households with "rural" characteristics.

Online Appendix Table 1: Causes of match failure

Category	Number
Potential matches: Unique by name/age, 1865	71,644
- Not unique by name/age in 1900 (*age misreporting*)	2,739
- Name/age combination not found in 1900	48,459
- *Due to death by 1900 (estimate)*	(18,200)
- *Due to name changes and transcription error (remainder)*	(30,259)
Actual matches: Unique by name/age, 1865 and 1900	20,446

Online Appendix Table 2: Comparing matched sample to the unmatched population by cause of match failure

	(1)	(2)	(3)	(4)	(5)	(6)
	Unique in 1865			Not unique in 1865		
	Unique 1900**	Not unique 1900	Not found 1900	Unique 1900	Not unique 1900	Not found 1900
Urban	0.259	0.292	0.252	0.119	0.090	0.124
	(0.438)	(0.427)	(0.434)	(0.324)	(0.286)	(0.330)
Urban Areas						
Household has assets	0.260	0.208	0.198	0.141	0.128	0.125
	(0.438)	(0.406)	(0.398)	(0.348)	(0.334)	(0.331)
Father occ. > median	0.591	0.538	0.537	0.471	0.458	0.437
	(0.491)	(0.499)	(0.498)	(0.499)	(0.498)	(0.496)
N	5,309	801	12,233	3,457	12,072	2,829
Rural Areas						
Household has assets	0.664	0.623	0.669	0.677	0.675	0.682
	(0.472)	(0.478)	(0.470)	(0.424)	(0.468)	(0.465)
Father occ. > median	0.607	0.549	0.598	0.573	0.546	0.573
	(0.432)	(0.497)	(0.490)	(0.494)	(0.497)	(0.494)
N	15,137	1,938	31,695	25,398	121,396	19,817

Notes: ** Matched sample

Online Appendix Table 3: Estimated earnings for farmers in the United States

	Statistics per farm
INCOME	
Value of farm products not fed to livestock	$753
Value of house rent and food/fuel produced on farm and consumed by family	$200 (*)
Gross earnings	***$953***
EXPENDITURES	
Labor, fertilizers	$98
Feed, seed, threshing	$75 (^)
Taxes	$27 (#)
Maintenance charges (building, machinery)	$62 (+)
Total	***$262***
NET EARNINGS	$691

Notes: (*) = Ratio of rent and food/fuel consumed to value of products sold from Goldenweiser (1916).
(^) = Ratio of feed, seed, and threshing charges relative to labor and fertilizers from Goldenweiser (1916).
(#) = Assume tax rate of 0.6% on total value of farm.
(+) = Assume maintenance charge (depreciation) of 0.05 on buildings and 0.15 on machinery. Values of buildings and machinery reported in 1900 Census of Agriculture.

Online Appendix Table 4: Estimated earnings for farmers in Norway

	Statistics per farmer
INCOME	
Value of farm products	$397 (+)
Value of house rent and food/fuel produced on farm and consumed by family (not reported)	$106 (*)
Gross earnings	$503
EXPENDITURES	$109 (*)
NET EARNINGS	$393

Notes: (+) = Unlike the US Census of Agriculture, the value of farm products is derived from transaction data, rather than farmer estimates. Therefore, we assume that the grain used on the farm to feed livestock is already excluded from the total.
(*) = We assume the same ratios as used for the US calculation.

Online Appendix Table 5: Estimated earnings for fisherman in Norway

	Statistics per fisherman
Value of products sold	$416
Share provided to fisherman	$145-$228. [We use $200.]
Value of direct consumption of fish	$48 (*)
TOTAL INCOME	$248

Notes: (*) Between 1830-1871, the average family spent 8 percent of their expenditures on fish (Grytten, 2004). The average Norwegian family's income was $300 (in US $1900), implying an expenditure of $24 on fish. The families of fisherman likely ate more fish than the average family. We double this value to $48.

Online Appendix Table 6: Comparing the matched sample with the Norwegian population in 1865

	Population	Match	Difference: (Match - Pop)
Urban	0.140	0.259	0.120
	(0.347)	(0.438)	(0.003)
A. Urban			
Name frequency	369.06	7.212	-361.84
	(1315.99)	(22.81)	(19.27)
Median	*11*	*2*	
Age	8.428	8.492	0.063
	(3.691)	(3.742)	(0.055)
Number of siblings	5.092	5.215	0.123
	(1.818)	(1.843)	(0.027)
Sibling rank	3.466	3.571	0.104
	(1.450)	(1.489)	(0.021)
Father occupation > median	0.498	0.592	0.104
	(0.500)	(0.492)	(0.022)
Household has assets	0.137	0.260	0.123
	(0.344)	(0.439)	(0.005)
B. Rural			
Name frequency	844.53	15.87	-829.45
	(2125.90)	(32.16)	(18.30)
Median	*121*	*7*	
Age	8.542	8.837	0.295
	(3.657)	(3.715)	(0.030)
Number of siblings	5.203	5.232	0.028
	(1.794)	(1.852)	(0.014)
Sibling rank	3.599	3.589	-0.009
	(1.498)	(1.492)	(0.012)

Father occupation > median	0.594 (0.491)	0.607 (0.487)	0.013 (0.004)
Household has assets	0.630 (0.483)	0.664 (0.473)	0.034 (0.004)
N	236,274	19,970	256,244

Notes: Column 1 contains means and standard deviations (in parentheses) of individual characteristics for the full population between the ages of 3 and 15 in Norway in 1865. Columns 2 reports similar statistics for the matched sample. Column 3 reports coefficients and standard errors for differences between the matched sample and the total Norwegian population. Name frequency counts the number of individuals in the full population (of all ages) with the same first and last name as the respondent. The number of siblings is inclusive of the individual. Oldest siblings have a sibling rank of one.

Online Appendix Table 7: Comparing the matched sample with the Norwegian population in 1900 and with Norwegian migrants in the United States in 1900

	Population	Match	Difference: (Match - Pop)
A. In Norway in 1900			
Age	43.811	43.824	0.011
	(3.725)	(3.724)	(0.030)
Married	0.864	0.855	-0.009
	(0.342)	(0.351)	(0.003)
Children	2.873	2.885	0.012
	(2.486)	(2.474)	(0.019)
ln(earnings)	5.773	5.814	0.040
	(0.425)	(0.451)	(0.004)
Urban residence	0.270	0.318	0.048
	(0.444)	(0.465)	(0.004)
N	139,535	17,432	156,967
B. In US in 1900			
Age	43.386	43.290	-0.095
	(3.734)	(3.755)	(0.164)
ln(earnings)	6.384	6.418	0.035
	(0.322)	(0.313)	(0.014)
Norwegian name index	1.432	1.478	0.046
	(0.471)	(0.435)	(0.014)
N	647	2,538	3,185

Notes: Column 1 contains means and standard deviations (in parentheses) of individual characteristics for the full population of men in Norway between the ages of 38 and 50 in 1900 and for men in this age range born in Norway and living in United States in 1900. The Norwegian data are taken from the full Census (100%), while the US data are drawn from the 1 percent IPUMS sample, thereby accounting for large differences in sample size. Column 2 reports similar statistics for the matched sample. Column 3 reports coefficients and standard errors for differences between the matched sample and the total Norwegian population. The Norwegian name index is equal to the sum of the probabilities that a man is born in Norway conditional on having a given first or last name; the index ranges from zero to two.