Summary of AER replication results [see AER replication tables for further details]

	Table 1. Common occupations held by Norwegian-born men	Figure 3. Occupational distribution of Norwegian-born men in the US and Norway.		Table 2. Return to migration from Norway to the US
Main result(s):	Migrants were more likely to be owner-occupier farmers and general laborers.	For born in rural areas, migrants hold somewhat lower paying jobs	For men born in urban areas, migrants more likely to hold low paying jobs.	Positive return to migration of approx. 57 to 64 log points
I. ABE algorithm				
NYSIIS names: exact age uniqueness band	✓	✓	✓	✓
NYSIIS names: 5-yr uniqueness band	✓	✓	✓	✓
NYSIIS names: 5-yr uniqueness band, exact age matches	✓	✓	✓	✓
Exact names: exact age uniqueness band	✓	somewhat positive	✓	✓
II. ABE using JW distance				
Using non-standardized names:				
Exact age uniqueness band	✓	✓	✓	✓
5-yr uniqueness band	✓	✓	✓	✓
Using Norwegian standardized names:		✓		
Exact age uniqueness band	✓	✓	✓	✓
5-yr uniqueness band	✓	✓	✓	✓
II. EM Algorithm				
Using non-standardized names:				
p = 0.80; l = 0.75	✓	✓	✓	✓
p = 0.90; l = 0.75	✓	✓	✓	✓
Using Norwegian standardized names:				
p = 0.80; l = 0.75	✓	somewhat positive	✓	✓
p = 0.90; l = 0.75	✓	✓	✓	✓

Main result(s):	Table 3 and Table 4. Direction of migrant selection from rural and urban areas.			
	Selection of urban migrants		**Selection of rural migrants**	
	Negative selection on household wealth (Table 4)	Implied negative selection on within-household unobservables (Table 3)	Negative selection on household wealth (Table 4)	Implied positive selection on within- household unobservables (Table 3)
I. ABE algorithm				
NYSIIS names: unique by exact age	✓	✓	mixed evidence*	✓
NYSIIS names: unique within 5-year band	✓	positive selection when weighted	mixed evidence*	✓
NYSIIS names: 5-year band, exact age matches	✓	positive selection when weighted	mixed evidence*	✓
Exact names: unique by exact age	mixed evidence*	✓	mixed evidence*	✓
III. ABE using JW distance				
Using non-standardized names:				
Unique by exact age	mixed evidence*	✓	✓	✓
Unique within 5-year band	✓	positive selection when weighted	mixed evidence*	✓
Using Norwegian standardized names:				
Unique by exact age	✓	✓	mixed evidence*	✓
Unique within 5-year band	mixed evidence*	✓	mixed evidence*	✓
V. EM Algorithm				
Using non-standardized names:				
p = 0.80; l = 0.75	✓	✓	✓	✓
p = 0.90; l = 0.75	✓	✓	mixed evidence*	✓
Using Norwegian standardized names:				
p = 0.80; l = 0.75	mixed evidence*	.	✓	.
p = 0.90; l = 0.75	mixed evidence*	.	✓	.

---

* Results show positive selection on some measures of father's wealth, but negative selection on others.

Memo on AER replication results

Summary

Our AER paper used US census data from Ancestry.com and the standard Ferrie/ABE matching algorithms (more and less conservative versions) to construct the panel data set for our analysis. Since our original study, there have been two main developments in historical record linkage. First, the full-count census data became available through the NBER. Second, a number of "second generation" matching algorithms have been developed. We first replicate the results of the original analysis using the newly-available full-count 1900 US census data with the original (ABE) matching algorithm. We then present results using "second-generation" matching algorithms such as the ABE algorithm with Jaro-Winkler adjustment (JW) and EM matching algorithms, and examine variations of the ABE algorithm that require an exact match on names and ages.

Here is a summary of how our key results replicate: We find that our measures of the returns to migration are generally similar across matching algorithms (Table 2). The finding that migrants from urban areas were negatively selected in their family's wealth is robust to most methods (Table 4). However, when we compare the OLS estimates of return to migration with within-household estimates in order to infer the direction of selection on unobservable characteristics, the finding that urban migrants were negatively selected on within-family characteristics is more sensitive to the matching method used (Table 3). We explain below how these results could be driven by different selection on family's wealth vs. other unobservables. The remainder of this document explains the replication steps and discusses in detail all replication results.

Replication steps and results

The specific matching methods we use in all our replications are outlined in detail here. One replication detail is especially relevant to this paper. In the US census, Norwegian names are frequently misspelled due to three main errors: 1) names with "ss" (i.e. Eriksson) are often spelled with a single "s", 2) vowels are often replaced (e.g., an "e" becomes an "a"), and 3) the letters "c" and "k" are often switched. To ensure that these spelling errors are not biasing the JW and EM samples, we also run a robustness check in which we first construct a "Norwegian standardization" of each name to account for these spelling errors, and then find matches using the JW and EM algorithms.

A note on sample sizes of brother pairs in Table3: the samples we use in our estimations are based on households for which we find migrant-stayer pairs. This allow us to estimate the returns to migration controlling for household fixed effects. An important difficulty in our replication exercise is that the number of pairs that the most conservative methods find is too small, especially for the weighted matched sample (because of missing information, primarily missing data on father's occupation). In this way, it is understandable that some of our results are more sensitive when using the most conservative methods. Let us give an example: the number of pairs we used for our baseline specification (method: ABE-NYSIIS) was 231 (respectively, 184 for our weighted sample). A more conservative version of this method (ABE-NYSIIS with a 5-year uniqueness band, and exact age matches) delivers only 20 pairs (and 13 for our weighted sample). Therefore, it does not come as a surprise that the results we get when we divide the sample in urban and rural areas (which are based on 6 or 7 pairs) tend to be more sensitive. In the most extreme case (when we use EM with Norwegian standardized names) we get 1 or 0 pairs. Therefore, we decided to omit this method from the analysis.

We report the key results from our replication exercises here:

- Table 1 and Figure 3 of the original paper summarize how the occupations of Norwegian-born men living in United States differ from those living in Norway. We originally found that migrants were more likely to work as laborers and owner-occupier farmers (Table 1) and that among urban men migrants were more likely to hold lower-ranked jobs than those who remained in Norway (Figure 3). These findings replicate well across all matching methods.

- We originally estimated that the return to migration was approximately 70 percent (Table 2). We continue to find similar estimates with all matching methods used in this replication. In the original paper, we found that return to migration was slightly lower when the more conservative ABE/Ferrie matching approach was used, and in our replication, we continue to find that the estimated return to migration is slightly lower when more restrictive versions of the ABE/Ferrie method and the EM algorithm are used.

- Table 3 of the original paper estimates the return to migration for only households with at least 2 sons in the matched panel data. We report estimates of the return to migration from OLS regressions and from regressions that include household fixed effects to eliminate unobserved household-level attributes correlated with both earnings and the propensity to migrate. In the original analysis, we found that the average return to migration in the fixed effects regressions was lower for migrants from urban households compared with rural households. Similar estimates of the return to migration are found across all matching methods, and in all cases rural migrants are found to have a higher return to migration[†]. Note that the number of matched migrant-stayer pairs included in the more conservative matched samples is quite low, and as such the results from these tables should be interpreted cautiously.

- In Table 3, we inferred the direction of migrant selection on household-specific characteristics by comparing the OLS and within-household estimates of return to migration. A larger within-household return to migration implies that the OLS estimate is downward biased by negative selection of migrant households, and a smaller within-household estimate implies positive selection. Table 4 instead provides evidence on the selection of migrants on family's wealth by comparing father's occupation and father's assets across migrant and non-migrant households. Note that Table 3 examines migrant selection along many unobserved household-level attributes, including household wealth, but also attributes such as social connections or ambition. Because father's wealth and other unobservable within-family characteristics may not influence migration in the same way, the direction of selection implied by Table 3 may not match the results found in Table 4.
  - Our original analysis found evidence of negative selection of urban migrants in both family's wealth (Table 4) and unobservable within-family characteristics (Table 3). When replicating these results with the full-count NBER census, we find even stronger evidence of negative selection on family's wealth from urban households, and we continue to find negative selection (although less significant) on unobservable within-family characteristics. Moreover, the finding that urban households are negatively selected on family's wealth is robust to most alternative matching methods.[‡] However, the finding of negative selection of urban household on within-family unobservable characteristics is sensitive to the matching method used (in some matched samples we find positive selection and in other we find negative).

[†] The one exception is the sample matched using the more conservative EM algorithm without standardized names, which finds higher returns among urban migrants. However, this table is based on very few observations.

[‡] The one exception is the sample matched using the more conservative EM algorithm without standardized names, which finds positive selection of migrants from urban areas. However, this table is based on very few observations.

o   Our original analysis found mixed evidence on the selection of migrants from rural areas. We originally found that rural migrants were negatively selected on family's wealth but positively selected on within-family unobservable characteristics. We continue to find mixed evidence on the selection of rural migrants, but the details are different. The evidence of positive selection of rural migrants on within-family unobservables characteristics (Table 3) is robust across all methods[§], however, the evidence of negative selection of rural migrants on family's wealth is sensitive to the matching method used, and in many cases, we find positive selection on some measures of family's wealth but negative selection on other measures.

[§] Again, the one exception is the sample matched using the more conservative EM algorithm without standardized names, which has a very small sample size and thus the results should be interpreted cautiously.