Long live your ancestors' American dream:
The self-selection and multigenerational mobility of American immigrants

Joakim Ruist*
University of Gothenburg
joakim.ruist@economics.gu.se

April 2017

Abstract

This paper aims to explain the high intergenerational persistence of inequality between groups of different ancestries in the US. Initial inequality between immigrant groups is interpreted as largely due to differently strong self-selection on unobservable skill endowments. These endowments are in turn assumed to be more persistent than observable outcomes across generations. If skill endowments are responsible for a larger share of total inequality between immigrant groups than between individuals generally, the former inequality will be more persistent. This explanation implies the additional testable hypothesis that the correlation between home country characteristics that influence the self-selection pattern – in particular the distance to the US – and migrants' or their descendants' outcomes will increase with every new generation of descendants. This prediction receives strong empirical support: The migration distance of those who moved to the US around the turn of the 20th century has risen from explaining only 14% of inequality between ancestry groups in the immigrant generation itself, to a full 49% in the generation of their great-grandchildren today.

Key words: migration; selection; intergenerational mobility; ancestry

JEL codes: F22, I24, J61, J62,

* I thank Mikael Lindahl and Jan Stuhler for very valuable comments and discussions.

1 – Introduction

It is well documented that ancestry is important for intergenerational mobility in the US. Inequality between groups of different ancestries[1] is more persistent from one generation to the next than inequality between individuals generally (Borjas, 1992, 1994). Candidate explanations for this pattern are few. The dominant hypothesis due to Borjas (1992) is that co-ethnics outside of the family, i.e. the "ethnic environment", have sizeable direct social impact on children's future outcomes in addition to that of the parents. Another potential explanation is discrimination. Yet to this date there exists no empirical evidence that speaks clearly in favor of one particular explanation.

This paper argues that to better understand the causes of the high persistence of inequality between Americans of different ancestries, we need to simultaneously consider the origins of this inequality; something that has previously been kept largely separate from analyses of mobility. Specifically we need to consider the ancestors' self-selection into migration. Migrants are generally strongly self-selected and far from random samples of the populations of their home societies. It is typically assumed that this self-selection is largely or predominantly on unobservable skill endowments (e.g. ability, preferences), and also that differences in self-selection patterns account for a large share of the inequality between immigrant groups from different countries (Chiswick, 1978; Borjas, 1987). The hypothesis of this paper is that this feature also explains the high intergenerational persistence of this inequality. The assumption that delivers this hypothesis is that the skill endowments on which migrants were selected are more strongly inherited – through nature or nurture – from one generation to the next than observable outcomes such as schooling or income are. Such strong intergenerational persistence of latent skill endowments is indicated e.g. by the recent empirical results of Clark (2014), and Braun and Stuhler (forthcoming). Due to self-selection, variation in observable outcomes between immigrant groups is more strongly correlated with variation in skill endowments than what is the case for outcome variation between most other groups or between individuals generally. Therefore, variation in observable outcomes between immigrant groups is also more persistent from one generation to the next, and the same is true between later generations of their descendants.

[1] These groups are commonly labelled "ethnic". Yet since the explanation I propose for why they are important in the intergenerational transmission process depends on ancestry but not on ethnic identification, I will instead refer to "ancestry groups".

This explanation for the high persistence of inequality between ancestry groups implies the additional testable hypothesis that home country characteristics that influence the strength of self-selection will be more and more strongly correlated with group-level outcomes with every new generation of descendants. This is because these characteristics are primarily correlated with the component of a group's socioeconomic situation that is due to skill endowments, and because this component declines more slowly than other components over the generations. The home country characteristic for which this prediction is evaluated empirically is the migration distance to the US. Among theoretically plausible candidates, it appears as the empirically strongest predictor of self-selection for recent migrants, and it is the only one for which measurability is not a problem for historical migrants. Theoretically, a longer migration distance implies a more positively self-selected migrant group, since the higher expected income gains from migration that come with higher endowments are necessary to cover the higher monetary and non-monetary costs that come with a longer migration distance.

The empirical support for this prediction is quite striking. In the sample of mostly fourth-generation immigrants observed in 2010-14, whose great-grandparents immigrated around the turn of the 20th century, the explanatory power of the great-grandparents' migration distance has risen to a full 49% of total inequality between ancestry groups, from only 14% among the great-grandparents themselves. Similar results are obtained when following a larger number of origins from a more recent cohort for two generations only. In the sample of children of immigrants observed in 2010, the migration distance of their parents explains 53% of total inequality between groups, up from 21% in the parent generation.

As an extension, I also evaluate the additional implicit prediction that the intergenerational persistence of inequality *within* a group of first-generation immigrants from the same country of origin should be particularly low. This prediction is the flip side of that of high persistence between these groups. If the sample that emigrates from a country is self-selected on having a certain level of skill endowments, then just like variation in these endowments constitutes a particularly large share of observable outcome variation between groups, it constitutes particularly small shares within them. Hence intergenerational mobility within these groups will be high from the first generation of immigrants to the second. In generations further away from the self-selected immigrant sample, endowment variance – and hence mobility – within groups will be higher. This prediction is also supported in the empirical analysis, although sample sizes are small and this result is also open for (at least) one alternative interpretation.

Section 2 of this paper provides the background and theoretical framework for the analysis. Section 3 describes the data and sample selections. The empirical analysis of inequality between ancestry groups is reported in Section 4, and that of inequality within groups in Section 5. Section 6 concludes.

2 – Setting and theory

The result that socioeconomic inequality between Americans of different ancestries is highly persistent across generations was first reported by Borjas (1992), who regressed outcome *y* (schooling or occupational prestige) of individual *i* of ancestry group *j* in generation *t* simultaneously on the same outcome of the individual's own father and the average outcome of the father's ancestry group in period *t-1*:

$$(1) \qquad y_{ijt} = \gamma_0 + \gamma_1 y_{ijt-1} + \gamma_2 y_{jt-1} + \xi_{ijt}$$

The estimates of the parameter γ_2 were consistently positive and quite large: between 0.10 and 0.46 across outcome variables and samples in the main analysis. For occupational prestige scores, the analysis even indicated that the ancestry group's average outcome had a larger influence on an individual of the next generation than that of the individual's own father. Borjas interpreted this result as a direct causal impact of co-ethnics outside of the family on children's future outcomes. This interpretation is commonly referred to as the "ethnic capital" hypothesis.

Later studies have confirmed this result in regressions at the ancestry group level. If intergenerational persistence is estimated at the group level, i.e. by the regression

$$y_{jt} = \delta_0 + \delta_1 y_{jt-1} + \vartheta_{jt}$$

the intergenerational coefficient obtained is the sum of the two at the individual level, i.e. $\delta_1 = \gamma_1 + \gamma_2$ (see Borjas, 1992: page 131). Hence the result that γ_2 is positive is equivalent to estimated persistence being higher at the ancestry group level than at the individual level: $\delta_1 > \gamma_1$. The latter result was reported by Borjas (1994), who estimated coefficients of persistence of log wage averages of ancestry groups as high as 0.6-0.7 from the first to the second generation of immigrants, and more uncertain yet only slightly lower coefficients from the second generation to the third. Somewhat lower but still high coefficients of persistence of log wages were also estimated by Borjas (1993), and Card, DiNardo, and Estes (2000).

However while suggesting a specific interpretation of this result, Borjas (1992) also noted that it is consistent with an importance of the ancestry group in general, such as e.g. due to discrimination or other.[2] This point was also recently made more formally by Braun and Stuhler (forthcoming) in the closely related context of estimating causal intergenerational “grandparent”, or “dynastic” effects, i.e. where the groups j in Equation (1) are extended families.[3] Also in this literature, positive estimates of the equivalent of γ_2 are commonly interpreted as reflecting direct causal impact of these extended family members (see e.g. Mare, 2011; Pfeffer, 2014), yet Braun and Stuhler point out that the result is consistent with “*any* causal process that generates sustained excess persistence”. In the context of ancestry groups, further empirical evidence that is more consistent with one such process than others does not exist to this date.

2.1 – The intergenerational mobility model

The hypothesis presented in this paper is that the high observed persistence of inequality between groups of different ancestries is due to the first generation of these groups, i.e. the immigrants, being differently strongly selected on unobservable skill endowments. Skill endowment variance therefore makes up a particularly large fraction of total outcome variance between groups. Skill endowments are in turn more persistent than observable outcomes across generations, making observed inequality particularly persistent when it is measured across ancestry groups.

Formalizing this, consider for simplicity families that consist of one individual only in each generation. Each individual i in generation t has skill endowments e, which are inherited according to the parameter λ:

$$e_{it} = \lambda e_{it-1} + k_{it}$$

The variable k is a random shock. Skill endowments are used in the production of human capital h:

$$h_{it} = \mu e_{it} + m_{it}$$

The skill endowment variable is expressed as deviations from its average. However to later account for differences in average human capital (yet not necessarily in its correlation with

[2] Others’ usage of the term “ethnic capital” has sometimes also included discrimination. See e.g. Solon, 2014.
[3] In most studies these are in practice grandparents only, yet e.g. Lindahl et al (2015) empirically investigate the wider extended family or “dynasty”.

skill endowments) across immigrants' countries of origin, average human capital needs to be explicit. Hence the part *m* of human capital that is orthogonal to endowments is the sum of the country's average level of human capital θ_j and a random shock *l*:

$$m_{it} = \theta_j + l_{it}$$

Also *m* is persistent across generations, as parents' human capital enhances the production of the human capital of their children. For simplicity keeping θ_j constant over time, the intergenerational transmission process for *m* is:

$$m_{it} = (1-\rho)\theta_j + \rho m_{it-1} + l_{it}$$

Hence *m* is inherited according to the parameter ρ, and multiplying θ_j by *(1-ρ)* implies the simplification that the variance of *m* is constant over time. Finally earnings *y* are a function of endowments and human capital:

$$y_{it} = \sigma e_{it} + \eta h_{it} + n_{it}$$

Where *n* is a random shock. This can in turn be written:

$$y_{it} = (\sigma + \mu\eta)e_{it} + \eta m_{it} + n_{it}$$

Where the expression inside the parenthesis gives the total return to skill endowments.

The intergenerational persistence (or "intergenerational elasticity") of human capital is given by β_h in the equation:

$$(2) \qquad h_{it} = \alpha_h + \beta_h h_{it-1} + e_{it}$$

Its probability limit is:

$$(3) \qquad plim\, \beta_h = \frac{\lambda\mu^2 Var(e_{it-1}) + \rho Var(m_{it-1})}{Var(h_{it-1})}$$

Substituting *y* for *h* in Equation (2) we get:

$$(4) \qquad plim\, \beta_y = \frac{\lambda(\sigma+\mu\eta)^2 Var(e_{it-1}) + \rho\eta^2 Var(m_{it-1})}{Var(y_{it-1})}$$

Hence the intergenerational persistence of an observable outcome is due to a combination of the two inheritance parameters λ and ρ, where the weights are the shares of total variance in the parent generation that are due to endowments and other inheritable components respectively. The crucial assumption for the predictions to be derived is that skill endowments are more persistent than observable outcomes across generations, i.e. $\lambda > \rho$. This hypothesis has recently been put forward and received some empirical support in Clark's (2014), Clark and Cummins' (2015), and Braun and Stuhler's (forthcoming) studies of multigenerational mobility (see also, and Stuhler, 2012, and Solon, 2015, for further discussions of the implications of different assumptions about transmission mechanisms). Yet while these studies raise the possibility that *all* intergenerational outcome persistence is due to skill endowment persistence, the multigenerational predictions and results of this paper also require a non-negligible degree of outcome persistence that is not so.[4] With multiple paths of intergenerational transmission, the model presented here thus most closely follows the approach of Conlisk (1969, 1974), and Nybom and Stuhler (2014).

Importantly, the probability limits in Equations (3) and (4) are valid regardless of at what level the analysis is conducted. Substituting between-group or within-group variances for total variances, they give the probability limits of estimated outcome persistence at the corresponding levels. Hence if observed persistence is different at different levels, this can be explained by differences in these variance shares.

2.2 – Immigrants' intergenerational mobility

The populations in all countries have the same endowment mean and variance. However they have different average education levels θ_j.[5] If immigrants were random samples of their home country populations, the model thus implies that intergenerational persistence would be *lower* between ancestry groups than between all individuals in a country. If selection was random, there would be zero endowment variation across ancestry group means. All outcome variation between groups in the immigrant generation ($t=1$) would be due to differences in average human capital between their countries of origin. Average human capital of immigrants from country *j* would simply be equal to θ_j, and average human capital in $t \geq 1$ would be:

$$h_{jt} = m_{jt} = \theta_{US} + \rho^{t-1}(\theta_j - \theta_{US})$$

[4] Here this additional persistence component is represented by the parameter ρ, and interpreted as the importance of parents' human capital in the production of their children's human capital. Similar but more algebraically complicated results can be obtained by instead modelling e.g. a feedback from parents' earnings to children's human capital, which would be interpreted as parents' monetary investments in their children's education.

[5] It is less certain though whether, and if so in what way, the correlation between endowments and human capital differs across countries. Hence μ is treated as constant across countries.

This process converges to θ_{US}, the average human capital level in the US, which for simplicity is treated as time-invariant. Without group-level variation in endowments, the between-group persistence rate of both human capital and earnings would be equal to ρ.

However immigrants are not random samples of their home country populations. They are strongly self-selected. We may expect them to be selected primarily either on their skill endowments or on their human capital. It is commonly assumed (e.g. Chiswick, 1978, Borjas, 1987) that unobservable endowments are central. This assumption is also crucial for the predictions of this paper. These do not require that self-selection be on endowments *only*; the component *m* of individual human capital may also play a role. However they do require that endowments are considerably more important than *m* in the self-selection process. Hence for expositional simplicity, I assume that only endowments matter.

If migrants are strongly enough self-selected on their skill endowments, the model presented here can explain why inequality between immigrant groups is more persistent across generations than inequality between native individuals. The strength of self-selection on endowments is different from different countries of origin, e.g. because of the variation across countries in returns to these skill endowments, and costs of migration to the US (see further in Section 2.4). Hence although countries' initial populations have identical endowment distributions, immigrant groups from these countries in the US do not. Average human capital of group *j* in *t≥1* is therefore:

$$(5) \qquad h_{jt} = \lambda^{t-1}\mu e_{j1} + \theta_{US} + \rho^{t-1}(\theta_j - \theta_{US})$$

Hence with large enough differences in the strength of self-selection across countries of origin, skill endowment variance makes up a larger fraction of total outcome variance between immigrant groups than between all individuals in the US. This in turn implies that estimated intergenerational mobility is lower between ancestry groups (see Equations (3) and (4)).

A closely related argument is made by Clark (2014). In Clark's model of intergenerational transmission, only endowments are inherited, i.e. similar to the present model with ρ*=0*. Clark argues that if the coefficient of intergenerational persistence is then estimated across groups with high within-group correlations in skill endowments (immigrant groups are once mentioned as plausible candidates among several others, yet are not in main focus), it will identify the "true" rate of intergenerational persistence, which is λ. Yet as Equations (3) and

(4) show, and as previously clarified by Clark and Cummins (2015), this is only true if *all* outcome variance between groups in the parent generation is due to skill endowment variance. Outcome variance due to other factors in the parent generation, inheritable or not, will bias the estimate downwards. By contrast, to explain the multigenerational results of the present paper, it is required that part of the outcome variation across ancestry groups is due to factors other than skill endowments, and furthermore that these too are transmitted across generations.

The aim of Clark (2014), and Clark and Cummins (2015) is to empirically estimate λ from the intergenerational persistence of inequality between groups that share rare surnames. Part of the criticism of this strategy by Chetty et al. (2014) is that these rare surnames are partly proxies for different ethnic groups, which implies that the strategy will pick up the high intergenerational persistence of inequality between these groups. Implicitly in their argument, this persistence is in turn due to factors other than skill endowments, implying that the assumption of zero group-level variance that is not due to skill endowments fails. Yet Chetty et al. acknowledge that little is known about the reasons for high persistence between ethnic groups, and conclude with a call for further investigation into this. The present paper aims to close this circle by arguing that the mechanisms of self-selection make it plausible that the reason for the high persistence of inequality between ethnic (ancestry) groups is indeed that this inequality is to a large extent due to variation in latent skill endowments, like Clark initially suggested was the case for groups based on surnames.[6]

Existing empirical evidence speaks neither in favor nor against the explanation for the low intergenerational mobility between immigrant groups presented here compared with e.g. ethnic capital or discrimination. However the explanation suggested here implies an additional testable prediction that I explore below.

2.3 – Selection pattern more visible in descendant generations

To evaluate the plausibility of the explanation proposed here, I will empirically evaluate an additional hypothesis that is implicit in the argument. If inequality between ancestry groups is to a particularly large extent due to variation in skill endowments, and if these endowments are more strongly inherited than observable outcomes, then skill endowments' share of total outcome inequality between ancestry groups will increase with every new generation of descendants of immigrants. Then if we can find a proxy variable to measure skill endowments, we will be able to observe this pattern.

[6] This argument has no implication for the appropriateness of estimation from groups based on surnames though.

In the immigrant generation, outcome variation between groups is the sum of one component that is due to endowment variation (because of self-selection), and one that is due to variation in education levels between home countries.[7] The first of these components will decline more slowly across generations, and apart from measurement error no additional group-level variation will be generated in descendant generations. The share of human capital variance in generation *t≥1* that is due to endowment variance in generation *t=1* is then:

$$\frac{\lambda^{2(t-1)}\mu^2 Var\left(e_{j1}\right)}{\lambda^{2(t-1)}\mu^2 Var\left(e_{j1}\right) + \rho^{2(t-1)} Var\left(\theta_j\right)}$$

This ratio increases with every new generation, as $\lambda > \rho$ implies that the numerator declines more slowly than the denominator. In principle this goes on indefinitely. Yet after a few generations there will be no discernable variation left in neither numerator nor denominator, as average human capital levels of all ancestry groups converge to θ_{US}.

Similarly, the share of earnings variance in generation *t≥1* that is due to endowment variance in generation *t=1* is:

$$\frac{\lambda^{2(t-1)}(\sigma + \mu\eta)^2 Var\left(e_{j1}\right)}{\lambda^{2(t-1)}(\sigma + \mu\eta)^2 Var\left(e_{j1}\right) + \eta^2\rho^{2(t-1)} Var\left(\theta_j\right)}$$

and increases over time for the same reason.

This prediction can also be expressed in a different way: The endowment average by group in *t=1* predicts absolute group-level mobility of subsequent generations, i.e. it is positively correlated with outcomes in *t* conditional on the same outcomes in *t-1*. If we could measure endowments directly we could estimate the regression equation:

$$h_{jt} = \beta_0 + \beta_1 h_{jt-1} + \beta_2 e_{j1} + \varepsilon_{jt}$$

We can write this as:

$$\lambda^{t-1}\mu e_{j1} + \rho^{t-1}\theta_j = \beta_1\left[\lambda^{t-2}\mu e_{j1} + \rho^{t-2}\theta_j\right] + \beta_2 e_{j1} + \varepsilon_{jt}$$

This would give us:

[7] This is of course a simplification. Other factors that may plausibly play important roles are e.g. limited transferability of human capital across countries, or outcome "luck" upon arrival in the US. However in the model these factors would play roles similar to that ascribed here to variation in human capital levels across home countries, and therefore this simplified interpretation does not affect the insights from the model.

$$plim\,\beta_1 = \rho$$

$$plim\,\beta_2 = \mu(\lambda - \rho)$$

Which are both positive. Substituting y for h we would get the same probability limit as above for β_1, and:

$$plim\,\beta_2 = (\sigma + \mu\eta)(\lambda - \rho)$$

Which is also positive.

Here we see why $\rho>0$ is required for the model's multigenerational predictions. If we set $\rho=0$ the predictions made here are only valid when comparing the first two generations, since already in the second generation all outcome variance between groups would be due to endowments. Yet with $0<\rho<\lambda$ convergence to the pattern predicted by endowments may go on for several generations.

The prediction that the correlation between endowments in $t=1$ and outcomes in $t\geq 1$ increases over time is not empirically useful in itself, since endowments cannot be observed. However, migrants' average skill endowments by country of origin will be correlated with country characteristics that influence the strength of self-selection. This implies the unusual situation that the unobservable can be observed by proxy. If we can identify an observable home country characteristic x_j that is correlated with e_{j1}, the predictions made here for e_{j1} will be valid also for x_j. Hence as above, the correlations between x_j and h_{jt} or y_{jt} will increase as t increases, and x_{jt} will predict h_{jt} and y_{jt} conditional on their lagged values.

2.4 – Predicting self-selection

To find candidates for x_j we need a theoretical model of migrants' self-selection. However it is clear from previous literature that such theoretical models are highly sensitive in the sense that small changes in unverifiable assumptions may strongly change the predictions obtained (e.g. Borjas, 1987; Chiswick, 1999; Grogger and Hanson, 2011). In this section I briefly present a basic theoretical framework, comment on the impact of changing some of its assumptions, and conclude that the search for an appropriate x_j variable should be predominantly an empirical question.

Following Sjaastad's (1962) seminal theoretical contribution, an income-maximizing individual of a certain type in a certain location will migrate to a different location if the implied discounted lifetime income increase is greater than the discounted lifetime monetary and non-monetary costs. Formalizing this, migration will happen if:

$$\Pi_{ikjd} = Y_{kd} - Y_{kj} - C_{jd} + t_{ikjd} > 0$$

Where *Π* denotes the net gain, *i* the individual, *k* the type, *j* the initial location, *d* the destination, *Y* the real income, *C* the migration costs, and *t* is the error term. All variables are discounted to net present values. It will be useful to write:

$$Y_{kj} = Y_j + R_{kj}$$

Where the income of type *k* in location *j* (similar in *d*) is the sum of the average income in *j* and the additional (positive or negative) return in *j* to being type *k*. We assume that the initial distribution across types is identical in all locations. The inflow of migrants into each destination *d* will then contain higher shares of those types whose returns are particularly high in *d*. Furthermore, it will do so most strongly for migrant flows from origins where returns to the same types are low, where average income is high, and from where the costs of moving to *d* are high. The first of these three is simply because the total outflow of type *k* is higher from where returns to *k* are lower. To see the latter two, we can collect all type-independent terms on one and all type-dependent terms on the other side of the inequality that determines when migration happens:

$$R_{kd} - R_{kj} > Y_j - Y_d + C_{jd} - t_{ikjd}$$

The higher country *j*'s average income or costs of moving to *d* on the right hand side of the inequality, the higher must the type-specific return be on the left hand side to make migration happen. Hence when the right hand side is large, only the types with the highest returns to living in *d* will migrate there.[8]

In the specific case at hand, our interest lies in the self-selection on skill endowments of migrants from different countries into the country where returns to these endowments are quite certainly the highest in the world, i.e. the US. The types in the model above are then defined by higher or lower skill endowment levels, and $R_{k,d=US} - R_{kj}$ is positive for all *j*. The model thus predicts that the migrant groups that originate in countries with low skill returns, high average income, and high costs of migration to the US will have the highest skill endowment levels.

[8] See Chiswick (1999) on the same point.

The basic theory presented here is of course not necessarily correct. There are several ways to change one assumption and arrive at markedly different predictions. For example, Borjas (1987) makes specific assumptions about the shape of the utility function and the correlation between skills and migration costs, and obtains the prediction that only the home country's relative skill returns R_{kj}/Y_j determine migrants' skill levels. Grogger and Hanson (2011) make a specific assumption about the distribution of the error term and obtain the prediction that R_{kj} and Y_j but not C_{jd} matter. Another option is to add a liquidity constraint

$$Y_j + R_{kj} \geq aC_{jd}$$

to the model,[9] where a is a positive scalar reflecting that discounting is different from in the previous equations. In this case the signs of the influences of R_{kj} and Y_j are ambiguous, depending on whether the constraint binds or not, while that of C_{jd} is still unambiguously positive.

It is therefore appropriate to view theoretical models as suitable for producing candidates for x_j, the country-level variable needed to test the empirical predictions of the previous subsection, rather than for excluding them before subjecting them to empirical testing. However, it should be noted beforehand that the three candidates identified here are highly different in terms of availability and quality of relevant data for historical migrant cohorts. The typical indicator of migration costs is the migration distance (e.g. Sjaastad, 1962; Schwartz, 1973). This variable has the considerable advantage of perfect data availability at the country level. It is also constant over time, implying that the question of at which point(s) in time a home-country variable is relevant for which migrants needs not being posed.

The quality of available measures of home country average income is poorer for historical migrants. Returns to latent skill endowments in a country are not possible to measure. The best available option (e.g. Borjas 1987, 1993) is probably to use information on income inequality, while assuming that the correlation between skills and earnings is the same in all countries. Yet the availability of good income inequality measures is severely limited already a few decades back from today. The migration distance between the home country and the US is thus preferable to the other x_j candidates for availability reasons. Fortunately, as will be seen in Section 4.1 (yet has received little attention in previous literature), it is also the

[9] Clark, Hatton, and Williamson (2007), and Hanson (2010) indicate that liquidity constraints are important in shaping international migration flows.

candidate that performs best in explaining migrant selection in recent years where availability is good also for the other candidates.

3 – Data and sample selections

The empirical analysis uses data from multiple years of censuses, ACS, and CPS. The data has been obtained through IPUMS (Ruggles et al., 2015). These data sets are the only ones that provide large enough numbers of individual observations within large enough numbers of ancestry groups to enable sufficient statistical power. They do not however provide any possibility of linking individual outcomes in one generation to the outcomes of these individuals' actual parents. Immigrants and their native-born descendants are thus, as in previous similar studies, linked by origin. Immigrant men from country *j* who are 25-60 years old in one year are considered the fathers of native-born individuals, with a father from country *j*, who are 25-60 years old approximately thirty years later. All analyses focus on the links between immigrant men and their male native-born descendants, to avoid contamination from differences in attitudes to female education and labor force participation across immigrant origins.

The main unit of analysis is the country of origin. Reported origins that are more specific (e.g. Sicily) are aggregated to countries. When countries have merged or split over time, typically some individuals report their origin in the larger aggregate while others do not. Hence consistency requires that the USSR, Czechoslovakia, and Yugoslavia are treated as merged units throughout. The exception from this rule will be Austria-Hungary, which ceased to exist before the sample period began, and the vast majority of respondents report their origin in either Austria or Hungary. In the analysis I ascribe the few that report Austria-Hungary to Austria, but changing this to Hungary has no discernible impact on the results.

To maximize both length and width, the analysis covers two different immigrant cohorts and their descendants. The late cohort consists of men who are 25-60 years old in 1980, and observed in the 5% sample of the census in that year. Their native-born children are observed in the CPS of 2005-14. The minimum requirement of a sample size of at least fifty individuals by origin is met by 107 countries of origin in the 1980 census, whereof by 52 also in the merged 2005-14 CPS. The larger ACS from the later period do not contain information on parents' place of birth and hence it is necessary to use the smaller CPS. Yet by merging ten survey years, a large enough sample is obtained. For simplicity, this merged sample is henceforth referred to as the year 2010. This cohort is included to maximize the width of the analysis, i.e. the number of origins. In 1980 the US had fairly large immigrant populations from substantially more countries of origin compared to one or two decades earlier. Yet 1980 is still early enough to enable observation of their native-born child generation in the same age interval thirty years later.

The early immigrant cohort consists of men who are 25-60 years old in 1930, and observed in the 100% sample of the census in that year. By choosing this year I can observe a maximum number of individuals from the great predominantly European immigration wave of around 1880-1930. This immigration peak can be seen in Figure 1, which shows US immigration by decade 1821-2010. In the 1930 cohort I can follow fewer countries of origin. Yet this lack of width is compensated by length: I can follow their descendants all the way up to a sample that on average contains their great-grandchildren, which I observe in 2010-14.

When estimating migrant selection models, I also include a third cohort that consists of men who are 25-60 years old in 2005-14 and observed in the ACS of these years. For simplicity, this merged sample is referred to as the year 2010. Although I cannot follow any descendants of this cohort, it is included in the selection analysis because of the substantially better data coverage of in particular income inequality measures in 2010 compared with 1980.

The outcome variables in the analysis of the 1980 and 2010 cohorts are average years of schooling and log weekly wages by ancestry. For the 1930 cohort, information on both these variables are lacking for the first cohort, i.e. in the 1930 census. Hence the analysis of this cohort primarily focuses on Hodge-Siegel-Rossi occupational prestige scores, which are available for all generations. However I also investigate results for years of schooling and log weekly wages of generations 2-4, where these are available.

The outcomes that are averaged by origin are the predicted outcomes from regression models on the entire samples of each year. For all samples, these regressions include a dummy for each age and US census division, and the predicted values refer to a 40-year-old who resides in the East North Central Division. Regressions in immigrant samples also include a dummy for each immigration year (intervalled in the 1980 census). Predicted values are for individuals who immigrated in 1915 for the 1930 sample, and in 1964-69 for the 1980 sample. Regressions in samples that merge several observation years include a dummy for each year, and predicted values refer to the center of the interval.

I also use information on characteristics of the migrants' home countries. The migration distance to the US is calculated as the distance in thousands of kilometers as the crow flies between the home country's capital and whichever of New York, Miami, and Los Angeles is closest. The average income in the home country is proxied by expenditure-side real GDP/capita at current PPP taken from the Penn World Tables version 9.0 (Feenstra et al., 2015). Income inequality in 2010 is measured as either the Gini coefficient or the income share held by the highest 20%; both variables from the World Bank's World Development Indicators. Data availability differs between the years; hence the 2010 values are averages of all available values for 2009-2011. Data on male average years of schooling in the home country is taken from Barro and Lee (2013).

3.1 – Identifying the third and fourth generations

Native-born men, with foreign-born fathers, who are 25-60 years old in 1960 and observed in the 5% sample of the census in that year are considered the sons of the 1930 immigrant cohort. To identify later descendants, like Borjas (1994) I use the Ancestry question of the census/ACS. Respondents are asked to name their "ancestry or ethnic origin". According to the census instructions, respondents who "have more than one origin and cannot identify with a single ancestry group may report two ancestry groups". In this case I assume that the ancestry that was noted first by the respondent is the most important one, and ascribe the individual to this ancestry. As a robustness check I also conduct all analyses including only individuals who reported one ancestry only.

The ancestry variable does not distinguish between first, second, and later generations of immigrants. For this purpose, separate information on own and parents' birthplaces is required. This is unproblematic for own birthplace, which is reported in all samples used. However no sample simultaneously contains information on ancestry and parents' birthplaces. Hence completely avoiding contamination from second-generation immigrants in the sample of third-generation immigrants, who are observed in the 5% sample of the 1990 census, is not possible. To minimize the problem, I use information from the 1995-98 CPS (the earliest years in which information on parents' birthplaces is available)[10] to estimate the sizes of the total US populations of second-generation immigrants by country of origin who were 25-60 years old in 1990. I use this information to exclude all origins where the share of second-

[10] Information on parents' birthplace is also available in the 1994 CPS, yet several of the countries in the sample are not separately coded in that year and therefore I do not use it.

generation immigrants in the native-born sample by ancestry in 1990 is thus estimated to be larger than one-fourth.[11]

Setting the limit to one-fourth is a natural choice given the distribution of the estimated shares. Among the 41 countries of origin that otherwise provide large enough samples in 1990 to be included in the analysis, the 22 lowest estimated shares of second-generation immigrants are quite uniformly distributed in the interval 0.0–0.21, from which there is a large discrete jump up to the 23rd lowest share at 0.33 and already the 27th is above one-half. The reason for this bimodal distribution is that the two periods of high immigration in American history, which were seen in Figure 1, were largely comprised of different origins. The first peak was predominantly European, but European immigration was much lower after 1930 and hence second-generation contamination in most samples of European origin in 1990 is low. Yet most non-European origins are strongly, in many cases almost exclusively, represented in the second peak and hence estimated contamination of the second generation in 1990 is high. Of the 22 countries with low enough contamination to be included in the sample, all except Japan, Lebanon, and Syria are European. I have further verified that the estimated shares of second-generation immigrants are not significantly correlated with any of the outcome variables in this sample.

The conclusion that the sample thus observed in 1990 consists of mainly third as opposed to later generations of immigrants is also drawn from a pattern that can be seen in Figure 1, i.e. that a very large share of pre-1950 immigration happened in 1880-1930. To enable an investigation into whether variation in the shares of later generations of immigrants in the third-generation sample correlate with average socioeconomic outcomes by origin in 1990, I first calculate the average immigration year of pre-1930 immigrants by country of origin. Since information on year of immigration was not collected in the censuses prior to 1900, I use the 1850, 1900, and 1930 censuses to calculate the average immigration year by country of origin using the formula:

$$imm\_year_j = \frac{1840 * N_{j1850} + av\_year_{j1900} * N_{j1900} + av\_year_{j1930} * N_{j1930}}{N_{j1850} + N_{j1900} + N_{j1930}}$$

Where N_{jyear} is the immigrant population from country j in *year*, and av_year_{jyear} is their average immigration year. For the year 1900 these are calculated only over immigrants who

[11] Since the CPS samples are small, I sample both females and males from both the CPS and the census when doing this.

arrived after 1850, and for 1930 over only those who arrived after 1900. Reflecting that immigration was low prior to 1830, the average immigration year of immigrants who are present in 1850 is assumed to be as late as 1840. This equation should give a fairly accurate estimate of the length of the average immigration history for all countries of origin except Britain, from where there was comparably large immigration also before 1800. Finally I have confirmed that this measure is not significantly correlated with any of the socioeconomic outcome measures in 1990.

The fourth-generation sample consists of men who are 25-60 years old when observed in the ACS of 2010-14 (henceforth 2012). The gap between the third and fourth generations is thus a bit short: only 22 years. Yet in relation to the first generation, which was observed in 1930, it implies an average generation length of 27 years between the first and fourth generations, which is probably even slightly better than the 30 years implied in the rest of the samples. The 2012 sample includes the same 22 countries of origin as the 1990 sample. Individuals are again ascribed to countries of origin based on their reported ancestry. Again, I rely on the immigration history pattern illustrated in Figure 1 to conclude that they are mainly immigrants of the fourth generation. I have also verified, using information on father's birthplace from the CPS of 2010-14, that estimated shares of second generation immigrants are low also in this sample.

3.2 – Intergenerational persistence of inequality between ancestry groups

A first illustration of the high intergenerational persistence of inequality between immigrant groups is given in Figure 2. The left panel correlates average log wages by origin of the first and second generations of the 1980 immigrant cohort. The slope of the regression line is 0.53 with a robust standard error of 0.15. The right panel does the same for occupational prestige scores of the first and fourth generations of the 1930 cohort. The slope of the regression line is 0.36 with a robust standard error of 0.11. Assuming an AR(1) process this implies a coefficient of persistence of $0.36^{1/3}$=0.71.[12]

A wider range of estimates of intergenerational persistence is reported in Table 1, with regression coefficients in column (1) and correlation coefficients in column (2). The outcome and generation pair used are indicated on the left of each row. Column (3) reports the coefficients of intergenerational persistence (β) implied by the regression estimates in (1)

[12] This value is reported for illustration. Note however that the theoretical model of this paper implies that the intergeneration process is not AR(1).

assuming AR(1) processes. For the 1930 cohort these estimates are all in the range 0.67-0.79. They are lower for the 1980 cohort, especially for the schooling variable. On the other hand the corresponding correlation coefficient is a full 0.82. In the first generation of this cohort, there is very large variation in schooling levels: the standard error across the 52 origins is 2.15 years, and the range is between 7.9 (Portugal) and 17.2 (India) years. In the next generation there is substantial convergence to the mean: the standard error falls to 0.9 years. Yet relative positions change little, as shown by the high correlation coefficient.

4 – Empirical analysis

In this section I evaluate the prediction that home country variables that are important in the migrant self-selection process will be more and more strongly correlated with a group's outcomes in the US with every new descendant generation. First I test the correlations between the candidate variables identified in Section 2.4 and outcomes of immigrants, to identify which candidate appears to be the best indicator of selection. I then proceed to testing the actual hypothesis.

4.1 – Migrant selection models

I estimate regression models where the dependent variable is average schooling or log wages of a migrant group in the US in 2010 or 1980, and the independent variables are the home country's distance to the US, log expenditure-side real GDP per capita, and Gini coefficient. I have also tried replacing the Gini coefficient with the income share held by the top 20%. This results in highly similar estimates, yet always with slightly higher p values. These results are not reported. Average years of schooling in the home country is included as a control variable in all regressions, to improve the interpretation of the coefficients on the other variables as measures of selection.

In Table 2 I report a large number of regression results, motivated by the fact that the importance of the migration distance for the skill content of international migration has previously not been given much attention. It is well-known that migration distance is a powerful predictor of the *size* of a bilateral migration flow (e.g. Clark, Hatton, and Williamson, 2007; Mayda 2010). Yet although the migration distance sometimes appears as a control variable in analyses of determinants of migrant groups' outcomes, its coefficients typically receive limited attention in spite of their often large predictive power (e.g. Borjas, 1993; Grogger and Hanson, 2011).[13]

In Panel A of Table 2 the dependent variable is average years of schooling among immigrants in the US in 2010. As an indication of the importance of self-selection of migrants, we may note that the coefficients on years of schooling in the home country are far below one: between 0.19 and 0.40 across the six specifications.[14] Turning to the selection proxy candidates, these enter the regressions separately in the first three columns. We see a strongly significant (T=6.9) coefficient with the expected positive sign on the migration distance, and also a significant (T=2.2) coefficient with the expected negative sign on the Gini coefficient. The coefficient on log GDP/capita is not significant. We may note that this variable is strongly correlated with years of schooling in the home country (the correlation coefficient is 0.78); hence possibly the sample size is too small to make the use of this variable as a measure of selection feasible. Columns (4)-(6) report the results from regressions where the independent variables of interest enter the regressions simultaneously. The coefficients on migration distance are still positive and strongly significant ($T \geq 5.3$). Yet those on the other two are not significant when the migration distance is also included in the regressions. Across columns, the magnitudes of the coefficients on distance are highly consistent, indicating that a distance increase by approximately 4,000 km implies one extra year of schooling among migrants.

Highly similar results are reported in Panel B, where the dependent variable is instead migrants' average log weekly wages in the US in 2010. Compared to Panel A there are some movements of the p values of the coefficients on log GDP/capita and the Gini coefficient around the 5% limit, and the results are similarly consistent in indicating a strong positive effect of the migration distance. A distance increase by 1,000 km is associated with around 3% higher wages.

Panel C reports similar results based on the sample of immigrants observed in the US in 1980, with schooling as the outcome variable in columns (1)-(3) and log wages in columns (4)-(6). These regressions do not include any inequality measure, due to poor coverage. The most striking difference from the 2010 results is probably that the coefficients on years of

[13] In fact Borjas (1993) even verified the additional prediction of this paper that migration distance positively explains not only the socioeconomic outcomes of immigrants but also the mobility of their children – although he did not elaborate on this result.

[14] An alternative interpretation is that immigrants have obtained education in the US. However the reported results change little if the sample is restricted to very recently arrived migrants.

schooling in the home country are even lower. There is no significant correlation between years of schooling in the home country and among migrants in 1980. Otherwise the results in Panel C are equally clear about the positive impact of the migration distance. Its coefficients are again strongly significant, whereas those on log GDP/capita are mostly not so. The magnitudes of the coefficients on distance are also highly similar to those that were estimated on the 2010 sample.

Taken together, the results reported in Table 2 give a strong and consistent indication that the migration distance to the US is the best proxy for migrant selection. Hence this is the variable I will use in the subsequent analysis. The migration distance is also the only home country variable that can be properly measured also in 1930; hence the strong performance of this variable in Table 2 is promising for the multigenerational analysis of the 1930 cohort. The 1930 cohort was not included in the results reported in Table 2, and it is not possible to control for home country schooling in that year. However between 82 countries of origin in 1930, an additional 1,000 km of migration distance implies a significant 0.33 points higher occupational prestige score (the robust standard error is 0.16).

4.2 – Migration distance and outcomes in later generations

An evaluation of the prediction that migration distance is more strongly correlated with outcomes in later generations is reported for the 1980 cohort in Figure 3. The left panel shows the correlation between distance and average schooling in the first generation. The correlation is strongly significant with $R^2=0.21$. However, as the right panel shows, the same correlation is far stronger in the generation of these migrants' children in 2010, where the parents' migration distance explains a full 53% of inequality between ancestry groups. The p value for the difference in R^2 between the first and second generations is below 0.001, based on 10,000 bootstrap replications. The corresponding results for log wages are not shown graphically, but show a similarly strong increase in R^2 from 0.12 in the first generation to 0.30 in the second.

A closer inspection of Figure 3 also reveals that the residuals from the linear regression lines included in the two graphs are strongly correlated. Their correlation coefficient is a full 0.80. Hence these residuals are not random noise around the regression line. Instead, as predicted, part of the residual from the first generation remains in the second (i.e. $\rho>0$), as the groups converge toward the pattern implied by their migration distances. This remaining part is approximately one-fourth, as a regression of the residuals of the second generation on those of the first gives a coefficient of 0.27.

The one ancestry group still really far off the prediction in the second generation is those of Laotian ancestry. This is not surprising. The Laotian immigrant group in 1980 consists almost entirely of refugees of war. Clearly a war lowers the utility cost of migration substantially, possibly even making it negative since the alternative cost may be death. Hence the high migration costs compared with other countries that are indicated by the long migration distance are likely exaggerated in the case of Laos (notably though, the same could be said of migrants from Cambodia, who are found close to those from Laos in the first generation, yet make a substantial upward movement in the second).

Figure 4 reports the corresponding pattern for the occupational prestige scores of the first two generations of the 1930 cohort. While the correlation between distance and prestige scores was significant in the full first-generation sample of 82 countries of origin, it is not so in the smaller sample of the 42 origins that also meet the sampling requirement of at least 50 observations in 1960. However in the second generation R^2 has risen from non-significant 0.05 in the first generation to significant 0.20. The difference in R^2 between the first and second generations is not significant though: its p value is 0.153, based on 10,000 bootstrap replications. Again the residuals from the two regressions are highly positively correlated with a correlation coefficient of 0.76. The one observation that is still far off its predicted value in the second generation represents the Philippines. Like for Laos in the previous figure, it is simple to argue that the migration distance overestimates migration costs from the Philippines to the US; in this case due to the close link between the countries in this period (in 1930 Filipino nationals were considered US citizens).

Finally Figure 5 reports the corresponding results for the first four generations of the smaller sample of groups that can be well enough identified also in the third and fourth generations (See section 3.1). It provides quite striking support for the prediction. In this now quite small sample, the correlations between distance and prestige scores are not statistically significant in either of the first two generations. Yet in the third it has become significant at the 1% level with R^2=0.39, and in the fourth generation the great-grandparents' migration distance explains a full 49% of total inequality between ancestry groups. The p value for the conclusion that R^2 is increasing on average – i.e. the p value from regressing the four R^2 values on a linear time trend – is 0.017, based on 10,000 bootstrap replications (the bootstrapped p value for the difference between the R^2 values of the first and fourth generations is 0.028). As in the previous analyses, the residuals from the linear regressions are strongly positively correlated across generations. Between the seven possible generation pairs the lowest correlation coefficient, which unsurprisingly is that between the first and fourth generations, is 0.54.

As was shown in Section 2.3, the prediction evaluated here also implies that migration distance should predict intergenerational mobility, i.e. that it should be positively correlated with outcomes in generation *t* conditional on those in *t-1*. The results from regressions that evaluate this prediction are reported in Table 3. The prediction is consistently supported. All coefficients in the table are positive and significant; those on the migration distance as well as those on parental outcomes.

However compared with those in Figures 3-5, the results in Table 3 are more open to alternative interpretations. In particular they may be interpreted as due to omitted characteristics in the parent generation. The variable that is used to measure outcomes in the parent generation does not give a full account of the socioeconomic status of that generation; if additional variables were included they would give a more complete picture together. If migration distance indeed influences selection, yet skill endowments are *not* more persistent than actual outcomes, positive coefficients on distance may still appear in Table 3 due to the correlation between distance and these omitted variables.[15] However this interpretation is reasonably only realistic for the first three columns, where we look at mobility between the first two generations. It appears quite unrealistic that it would explain also the mobility of later generations. This type of mechanism would also not create the pattern reported in Figures 3-5; R^2 would still decline over time.[16]

All results reported in this subsection on migration distance and outcomes by generation look highly similar if the third and fourth generation samples are restricted to individuals who report only one ancestry. For the analyses of native-born descendant generations of the 1930 cohort they also look highly similar for schooling or log wage outcomes as for prestige scores. These results are therefore not reported.

5 – Extension: Mobility within ancestry groups

[15] Similarly, as noted by Borjas (1992, pages 141-142), it is theoretically possible – although perhaps not plausible – that omitted characteristics in the parent generation could explain the high persistence of group-level inequality as such.

[16] Beyond these observations, much more cannot be done to account for this alternative interpretation, due to the nature of the data. As Warren and Hauser (1997), and Braun and Stuhler (forthcoming) have shown conceptually in a similar context of estimating conditional grandparent effects, if this kind of omitted outcome bias is present the coefficient on distance should decrease as we control for additional outcome measures in the parent generation. However exploring this is data-demanding and not suitable for small samples like the present.

In section I evaluate empirically an additional implication of the theoretical model presented in this paper, namely that the persistence of inequality *within* ancestry groups is lower from the immigrant generation to their children than between later generations. This is again because of differences in variance shares. Migrants from country *j* are a subsample of the total population of country *j*, which is defined by having a certain level of endowments. Hence $Var(e_{it})$ will be smaller within the sample of immigrants from *j* in the US than in the countries' total populations (pre-migration it is the same in all countries). However, $Var(m_{it})$ will be the same.[17] Hence according to Equations (3) and (4), within the group from country *j* persistence will be particularly low from the migrant generation to their children. It will then approach the total population rate of persistence as $Var(e_{it})$ increases with every new generation, while $Var(m_{it})$ stays constant. The implication of this argument for the *total* (i.e. the sum of between- and within-group) mobility of immigrants versus natives is ambiguous though. It depends – again according to Equations (3) and (4) – on the variance shares that are between and within groups respectively.

Notably though, also other mechanisms could generate comparatively high intergenerational mobility within immigrant groups. It is possible to imagine e.g. that foreign-born parents have a weaker role model impact on their native-born children than native-born parents do, simply because they are foreign-born. Yet models that explain the high intergenerational persistence of inequality between ancestry groups by ethnic capital or discrimination do not imply any predictions regarding mobility within these groups.

Previous empirical studies do not provide any evaluation of this hypothesis. Borjas (1992) reports coefficients of intergenerational persistence within ancestry groups, but does not distinguish between generations with foreign- and native-born parents in this part of the analysis.

To evaluate this prediction empirically, it is necessary to use data where socioeconomic outcomes can be linked between individual men and their fathers. Hence the census/CPS/ACS cannot be used. Among data sets providing this possibility, the General Social Surveys (GSS) are chosen, because they provide the most detailed coding of ancestry. The samples of native-born men with foreign-born fathers are small though; hence I merge data from the 2002-14 biannual waves to obtain a large enough sample. The sample consists of all men aged 25 or

[17] This point is easily generalized. If *m* also matters in the self-selection process, also $Var(m_{it})$ will be lower within the migrant group; yet not so to the same extent as $Var(e_{it})$, as long as *e* matters substantially more than *m* for self-selection.

older. The schooling variables measure the numbers of years of schooling of respondents and their fathers respectively. No information on earnings is available. Hence I use the Hodge-Siegel-Rossi occupational prestige scores, which are available for both respondents and their fathers. Ancestry is measured by the survey question on country of family origin.

To test the hypothesis of higher mobility within ancestry groups for native-born men when the father is foreign-born, I run regressions on the form of Equation (2) separately for native-born men with native- and foreign-born fathers respectively, while adding a dummy for each ancestry to make the estimated coefficients of persistence refer to persistence within ancestry groups. The results are reported in Table 4. The first column reports a within-group intergenerational schooling elasticity of 0.18 in the sample with foreign-born fathers, and the second a corresponding elasticity of 0.29 in the sample with native-born fathers. As predicted, the latter coefficient is larger, and the difference between them is significant at the 1% level. Although the elasticity is lower, R^2 is substantially higher in the sample with immigrant fathers, because – as expected – the ancestry fixed effects explain a larger share of the variation in this sample.

Columns (3) and (4) report the results from a similar comparison of intergenerational prestige score elasticities. Here the difference between point estimates is even larger than for schooling: 0.10 versus 0.24. It is not statistically significant though. The sample of native-born individuals with foreign-born fathers in column (1) was already small: 283 observations. Yet the subsample of these who work and report their occupation is even smaller, 176 observations, resulting in a large standard error in column (3).

The small sample size also prevents a more elaborate analysis beyond the simple comparisons reported here. We may conclude that both results reported are in line with the predictions, and for the part of the analysis that was based on the larger sample this result is highly significant.

6 – Conclusion

In this paper I have suggested a theoretical model of migrant selection and intergenerational mobility that offers a potential explanation for the high persistence of inequality between Americans of different ancestries. Migrants are self-selected on their unobservable skill endowments, and these endowments are more persistent than observable outcomes between generations. I have supported this explanation by empirically verifying its implicit prediction that the correlation between a home country characteristic that influences the strength of self-selection – here the migration distance – and groups' socioeconomic outcomes in the US increases with every new generation of descendants of migrants.

The policy relevance of this result lies to a large extent in what it does not say. It is well-known that inequality between ancestry groups in America is highly persistent, and also that some groups experience more mobility than others. Previous explanations for this to some extent indicate that something is "wrong", in that certain groups' upward mobility is hampered either by these groups' own behavior, or American society's behavior towards them. As such they also indicate a role for policy in improving the situation. In contrast, according to the results and interpretation reported here, ancestry groups' low socioeconomic mobility is not an indication that something is wrong, but merely that the impact of migrants' self-selection is longer-lasting than previously thought.

The results also lend support to a particular type of intergenerational transmission model. The model that underlies this study follows in the recent tradition of models that assume that latent unobservable endowments are the most important or most persistent component in intergenerational transmission. Building on similar assumptions, Clark (2014) has argued that intergenerational mobility is overestimated in all countries, and Braun and Stuhler (forthcoming) have argued for a non-causal interpretation of the positive correlation between individuals' and their grandparents' (or extended families') socioeconomic outcomes conditional on those of their parents. This study adds to the empirical evidence in favor of this type of intergenerational model. However it does not support the strongest form of this model, where latent endowments are the *only* factor that is inherited from one generation to the next. The multigenerational results reported both on asymmetric assimilation and on increasing correlation with migration distance in later generations indicate that actual outcomes are inherited too, although to a lesser degree than endowments.

The results also support the common assumption in the literature on migrant self-selection that unobserved skills are central in the selection process. This assumption is generally not testable, since a correlation between e.g. migrants' schooling levels and a home-country variable that should theoretically influence their self-selection is generally not informative on whether selection is on schooling itself, or on a variable such as latent skills that also influences how much schooling is obtained prior to migration. By contrast the results and interpretation reported in this paper specifically require that latent endowments are more important than obtained schooling in the selection process.

Finally these results say something important not only about migrants' self-selection and intergenerational mobility, but also about America. In the 19th century, many millions of Europeans dreamed of a new life in America. But the journey was costly, and at least until the arrival of transatlantic steamships even dangerous, and only some actually made the leap. The results reported in this study not only support the view that those who actually did make the journey were equipped with qualities not equally possessed by all of those who did not. They also tell us that these qualities remained for several generations with their descendants, who made their native country the global hub of knowledge, innovation, entrepreneurship, and industry of the 20th century.

References

Barro, Robert, and Jong-Wha Lee (2013), "A new data set of educational attainment in the world 1950-2010", *Journal of Development Economics*, 104: 184-198

Borjas, George (1987), "Self-selection and the earnings of immigrants", *American Economic Review*, 77: 531-553

Borjas, George (1992), "Ethnic capital and intergenerational mobility", *Quarterly Journal of Economics*, 107: 123-150

Borjas, George (1993), "The intergenerational mobility of immigrants", *Journal of Labor Economics*, 11: 113-135

Borjas, George (1994), "Long-run convergence of ethnic skill differentials: the children and grandchildren of the great migration", *Industrial and Labor Relations Review*, 47: 553-573

Braun, Sebastian, and Jan Stuhler (forthcoming), "The transmission of inequality across multiple generations: testing recent theories with evidence from Germany", *The Economic Journal*, forthcoming

Card, David, John DiNardo, and Eugena Estes (2000), "The more things change: immigrants and the children of immigrants in the 1940s, the 1970s, and the 1990s", in George Borjas (ed), I*ssues in the Economics of Immigration*, University of Chicago Press, 227-269

Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez (2014), "Where is the land of opportunity? The geography of intergenerational mobility in the United States", *Quarterly Journal of Economics*, 129: 1553-1623 (online appendix)

Chiswick, Barry (1978), "The effect of Americanization on the earnings of foreign-born men", *Journal of Political Economy*, 86: 897-921

Chiswick, Barry (1999), "Are immigrants favorably self-selected?", *American Economic Review, AEA Papers and Proceedings*, 89: 181-185

Clark, Gregory (2014), *The son also rises*, Princeton University Press

Clark, Gregory, and Neil Cummins (2015), "Intergenerational wealth mobility in England 1858-2012: Surnames and social mobility", *The Economic Journal*, 125: 61-85

Clark, Ximena, Timothy Hatton, and Jeffrey Williamson (2007), "Explaining U.S. immigration 1971-1998", *Review of Economics and Statistics*, 89: 359-373

Conlisk, John (1969), "An approach to the theory of inequality in the size distribution of income", *Economic Inquiry*, 7: 180-186

Conlisk, John (1974), "Can equalization of opportunity reduce social mobility?", *American Economic Review*, 64: 80-90

Feenstra, Robert, Robert Inklaar, and Marcel Timmer (2015), "The next generation of the Penn World Table", *American Economic Review*, 105: 3150-3182

Grogger, Jeffrey, and Gordon Hanson (2011), "Income maximization and the selection and sorting of international migrants", *Journal of Development Economics*, 95: 42-57

Hanson, Gordon (2010), "International migration and development", in Ravi Kanbur and Michael Spence (eds), *Equity and growth in a globalizing world*, The World Bank, pages 229-262

Lindahl, Mikael, Mårten Palme, Sofia Sandgren Massih, and Anna Sjögren (2015), "Long-term intergenerational persistence of human capital: an empirical analysis of four generations", *Journal of Human Resources*, 50: 1-33

Mare, Robert (2011), "A multigenerational view of inequality", *Demography*, 48: 1-23

Mayda, Anna Maria (2010), "International migration: a panel data analysis of the determinants of bilateral flows", *Journal of Population Economics*, 23: 1249-1274

Nybom, Martin, and Jan Stuhler (2014), *Interpreting trends in intergenerational mobility*, Swedish Institute for Social Research Working Paper 3/2014

Pfeffer, Fabian (2014), "Multigenerational approaches to social mobility. A multifaceted research agenda", *Research in Social Stratification and Mobility*, 35: 1-12

Ruggles, Steven, Katie Genadek, Ronald Goeken, Josiah Grover, and Matthew Sobek (2015), *Integrated Public Use Microdata Series: Version 6.0* [Machine-readable database], University of Minnesota

Schwartz, Ada (1973), "Interpreting the effect of distance on migration", *Journal of Political Economy*, 81: 1153-1169

Sjaastad, Larry (1962), "The costs and returns of human migration", *Journal of Political Economy*, 70: 80-93

Solon, Gary (2014), "Theoretical models of inequality transmission across multiple generations", *Research in Social Stratification and Mobility*, 35: 13-18

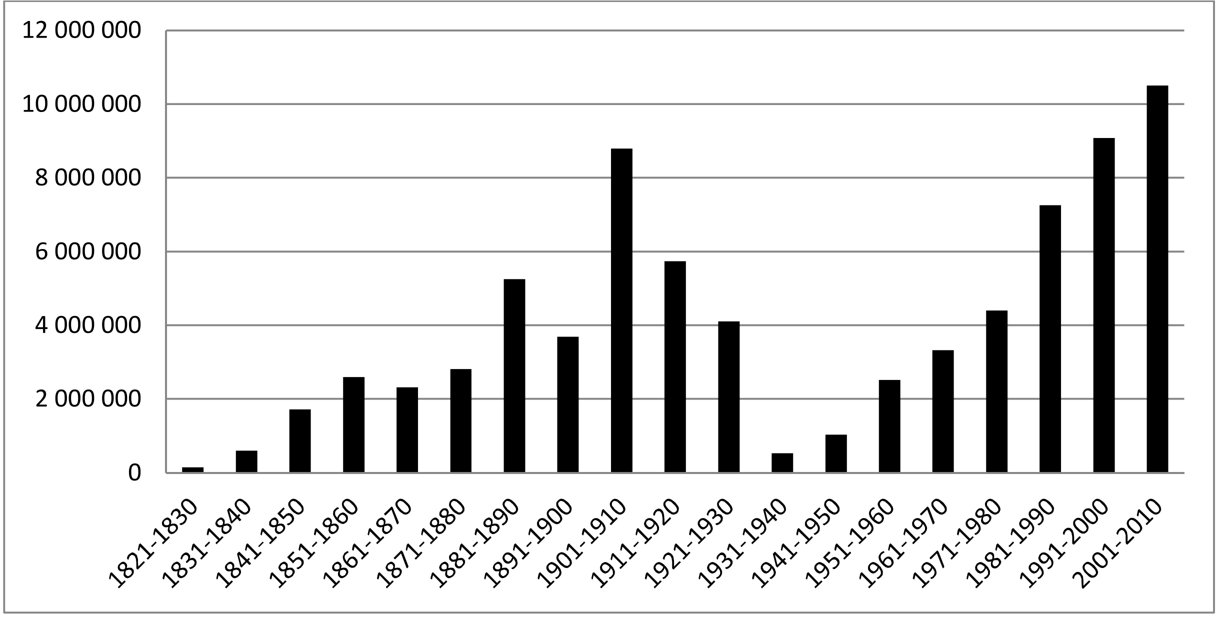
Solon, Gary (2015), "What do we know so far about multigenerational mobility?", *The Economic Journal*, forthcoming

Stuhler, Jan (2012), *Mobility across multiple generations: the iterated regression fallacy*, IZA Discussion Paper No. 7072

Warren, John Robert, and Robert Hauser (1997), "Social stratification across three generations: new evidence from the Wisconsin Longitudinal Study", *American Sociological Review*, 62: 561-572

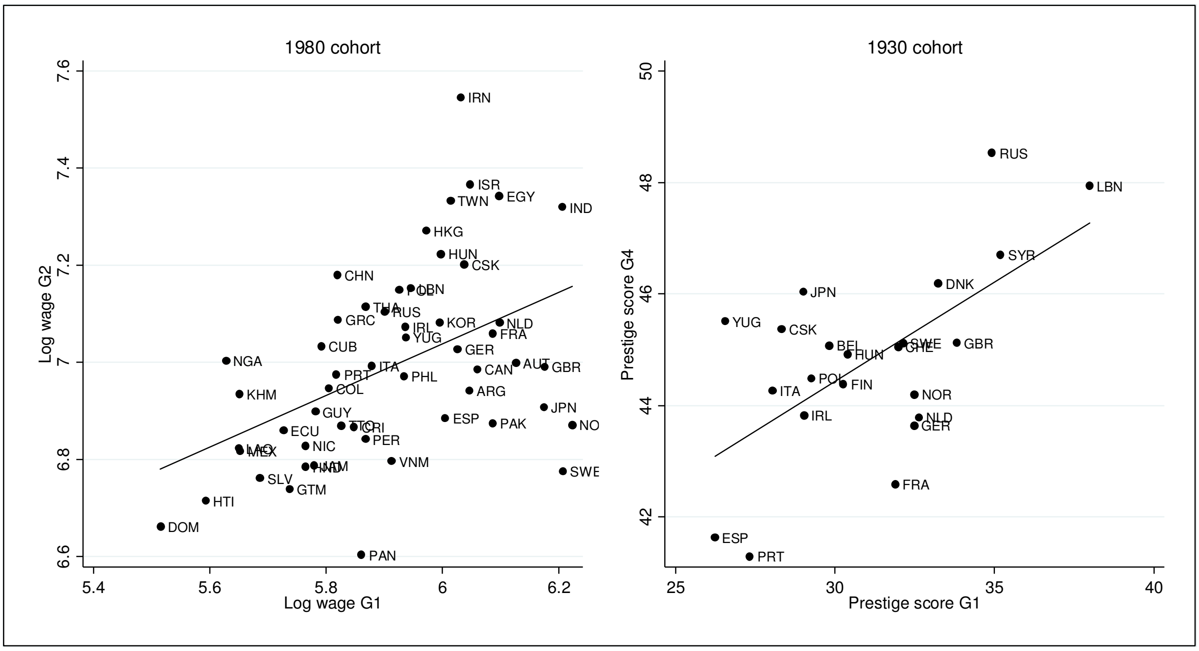
Figures

Figure 1. US immigration by decade 1821-2010



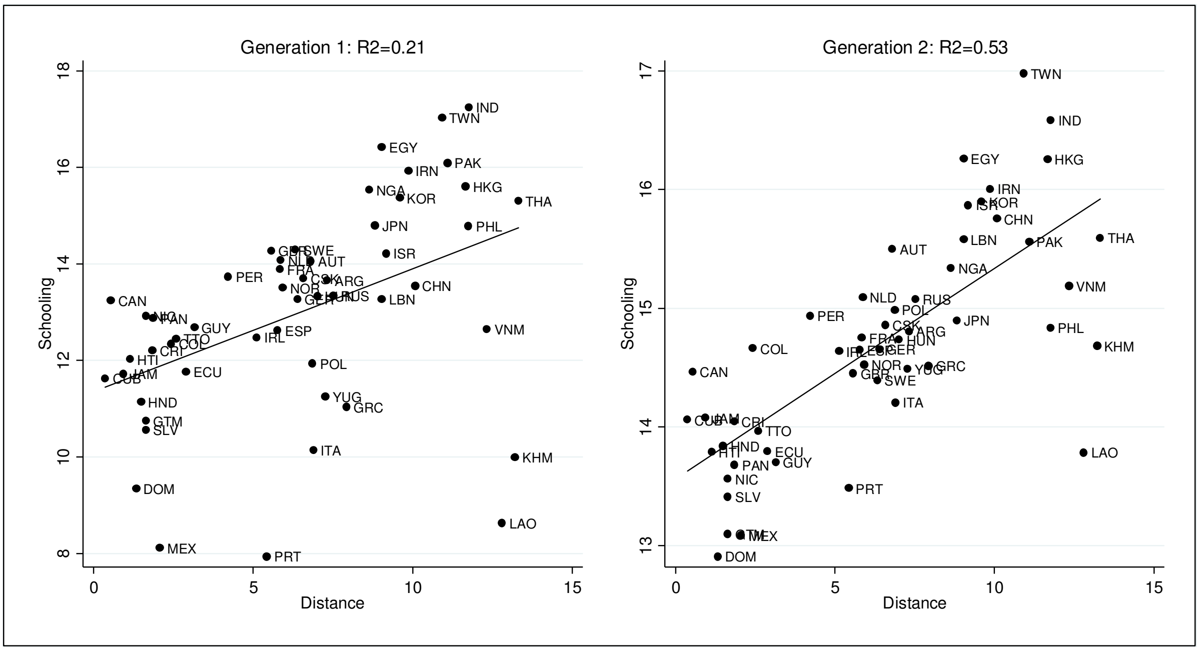
Notes: Data source: 2014 Yearbook of Immigration Statistics, US Department of Homeland Security

Figure 2. Intergenerational persistence of immigrants’ socioeconomic outcomes



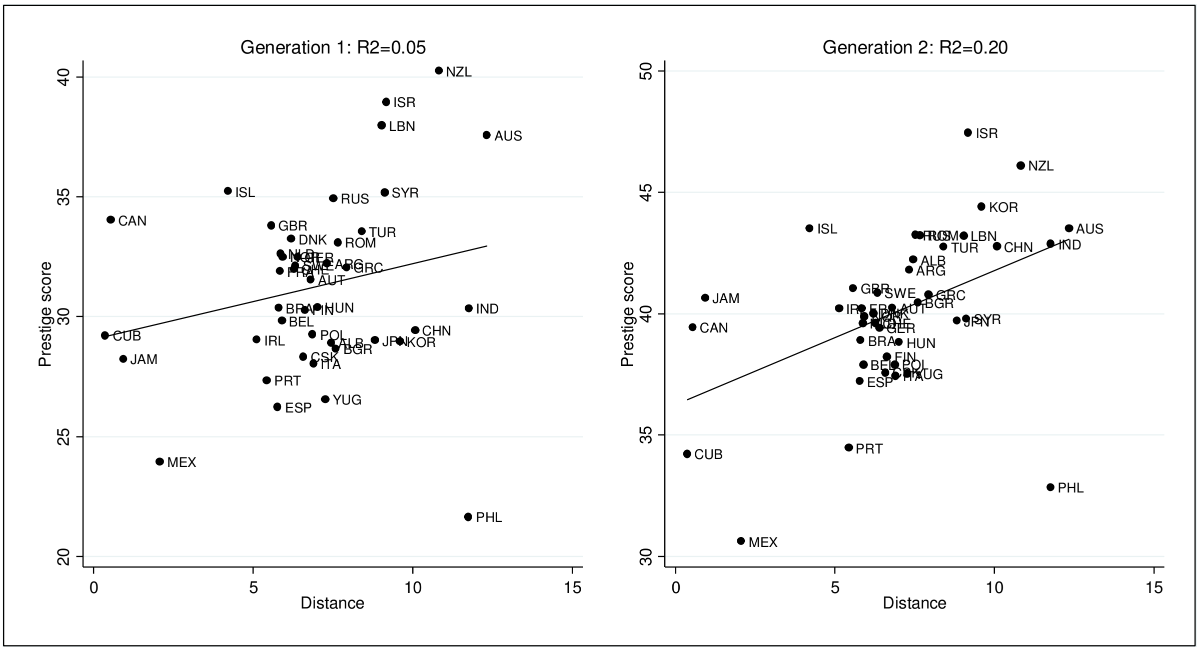
Notes: The graph to the left (n=52) shows average log wages of immigrants in 1980 (G1), and of their children in 2010 (G2). The graph to the right (n=22) shows average occupational prestige scores of immigrants in 1930 (G1), and of their great-grandchildren in 2012 (G4). Each observation is a migrant group from one country of origin in the US.

Figure 3. Migration distance and average schooling: 1980 cohort generations 1-2



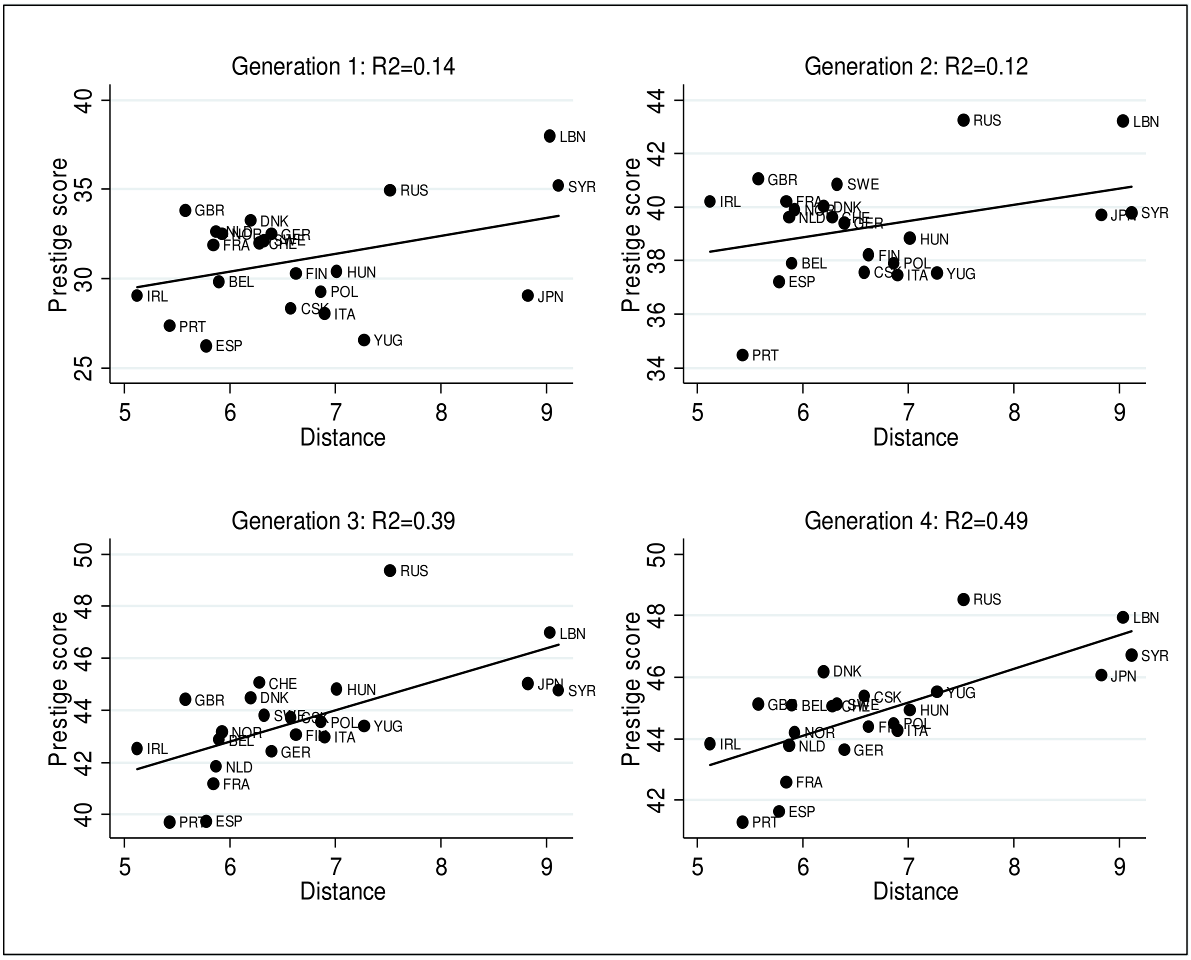
Notes: N=52. Each observation is a migrant group from one country of origin in the US. Distance is measured in thousands of kilometers.

Figure 4: Migration distance and average prestige score: 1930 cohort generations 1-2



Notes: N=42. Each observation is a migrant group from one country of origin in the US. Distance is measured in thousands of kilometers.

Figure 5. Migration distance and average prestige score: 1930 cohort generations 1-4



Notes: N=22. Each observation is a migrant group from one country of origin in the US. Distance is measured in thousands of kilometers.

Tables

Table 1. Intergenerational persistence between ancestry groups

	(1) Regression coefficient	(2) Correlation coefficient	(3) β implied by (1) assuming AR(1)	(4) Observations
1980 Cohort				
--- G1-G2 Schooling	0.355** (0.035)	0.821	0.355	52
--- G1-G2 Log wage	0.531** (0.147)	0.469	0.531	52
1930 Cohort				
--- G1-G2 Prestige score	0.671** (0.087)	0.767	0.671	42
--- G1-G3 Prestige score	0.446** (0.128)	0.627	0.668	22
--- G1-G4 Prestige score	0.357** (0.109)	0.617	0.710	22
--- G2-G3 Prestige score	0.794** (0.163)	0.734	0.794	22
--- G2-G4 Prestige score	0.606** (0.124)	0.688	0.778	22
--- G3-G4 Prestige score	0.772** (0.062)	0.949	0.772	22

Notes: Each row represents the estimated coefficient of intergenerational persistence using the outcome and generation pair indicated to the left. The β estimates in (3) are calculated from (1) assuming an AR(1) process. Robust standard errors in parentheses. * = $p<0.05$, **=$p<0.01$.

Table 2. Migrant selection models

	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A					
	2010 – Schooling					
Years of schooling	0.314**	0.333**	0.189*	0.348**	0.277**	0.396**
	(0.053)	(0.119)	(0.072)	(0.097)	(0.065)	(0.135)
Distance	0.238**			0.238**	0.259**	0.245**
	(0.035)			(0.036)	(0.045)	(0.046)
Log GDP/capita		-0.193		-0.0706		-0.336
		(0.278)		(0.205)		(0.359)
Gini			-0.0612*		-0.0328	-0.0359
			(0.0282)		(0.0179)	(0.0181)
R2	0.46	0.18	0.25	0.47	0.54	0.57
N	100	94	67	94	67	65
	Panel B					
	2010 – Log wage					
Years of schooling	0.0729**	0.0505**	0.0617**	0.0524**	0.0707**	0.0552**
	(0.0071)	(0.0127)	(0.0102)	(0.0102)	(0.0101)	(0.0178)
Distance	0.0304**			0.0312**	0.0266**	0.0307**
	(0.0044)			(0.0046)	(0.0075)	(0.0081)
Log GDP/capita		0.0519		0.0679**		0.0569
		(0.0333)		(0.0235)		(0.0489)
Gini			-0.00795		-0.00503	-0.00432
			(0.00491)		(0.00413)	(0.00413)
R2	0.54	0.41	0.47	0.59	0.58	0.62
N	100	94	67	94	67	65
	Panel C					
	1980 – Schooling			1980 – Log wage		
Years of schooling	0.0691	0.142	0.0738	0.0277**	0.0169	0.0120
	(0.0628)	(0.128)	(0.0980)	(0.0049)	(0.0091)	(0.0093)
Distance	0.276**		0.300**	0.0198**		0.0216**
	(0.047)		(0.045)	(0.0041)		(0.0046)
Log GDP/capita		-0.402	0.019		0.0273	0.0576*
		(0.370)	(0.309)		(0.0248)	(0.0279)
R2	0.27	0.02	0.29	0.31	0.15	0.36
N	96	83	83	96	83	83

Notes: Each observation is a migrant group from one country of origin in the US. The dependent variable and year of observation is indicated in the panel head. Distance is measured in thousands of kilometers. Robust standard errors in parentheses. * = p<0.05, **=p<0.01.

Table 3. Migration distance and mobility

	(1) 1980 cohort G2 Schooling	(2) 1980 cohort G2 Log wage	(3) 1930 cohort G2 Prestige score	(4) 1930 cohort G3 Prestige score	(5) 1930 cohort G4 Prestige score
Distance	0.109**	0.0226**	0.358*	0.803**	0.280*
	(0.016)	(0.0065)	(0.162)	(0.215)	(0.108)
Schooling(t-1)	0.267**				
	(0.026)				
Log wage(t-1)		0.362*			
		(0.140)			
Prestige score(t-1)			0.614**	0.636**	0.681**
			(0.098)	(0.163)	(0.064)
R2	0.83	0.39	0.67	0.69	0.92
N	52	52	42	22	22

Notes: Each observation is a migrant group from one country of origin in the US. The dependent variable, cohort, and generation are indicated in the column heads. Distance is measured in thousands of kilometers. Robust standard errors in parentheses. * = p<0.05, **=p<0.01.

Table 4. Estimated intergenerational elasticities between fathers and native-born sons

	Schooling		Prestige scores	
	(1)	(2)	(3)	(4)
	Immigrant fathers	Native fathers	Immigrant fathers	Native fathers
Father schooling	0.179**	0.288**		
	(0.036)	(0.012)		
Father prestige score			0.100	0.236**
			(0.092)	(0.022)
R2	0.34	0.20	0.29	0.10
N	283	3,384	176	2,384

Notes: Each coefficient is the intergenerational elasticity between fathers and their sons. All regressions include ancestry and year fixed effects. * = p<0.05, **=p<0.01.