ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Ruist, Joakim

Working Paper

Long live the American dream: Self-selection and inequality-persistence among American immigrants

CReAM Discussion Paper Series, No. 14/17

Provided in Cooperation with:
Rockwool Foundation Berlin (RF Berlin)

Suggested Citation: Ruist, Joakim (2017) : Long live the American dream: Self-selection and inequality-persistence among American immigrants, CReAM Discussion Paper Series, No. 14/17, Centre for Research & Analysis of Migration (CReAM), Department of Economics, University College London, London

This Version is available at:
https://hdl.handle.net/10419/295571

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

CReAM
Centre for Research & Analysis of Migration

Discussion Paper Series

CPD 14/17

- **Long live the American dream: Self-selection and inequality-persistence among American**
- Joakim Ruist

Centre for Research and Analysis of Migration
Department of Economics, University College London
Drayton House, 30 Gordon Street, London WC1H 0AX

www.cream-migration.org

Long live the American dream:
Self-selection and inequality-persistence among American immigrants

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

Abstract

This paper aims to explain the slow economic convergence between groups of different ancestries in the US, i.e. why these groups experience even less intergenerational mobility than individuals in the same country. It shows how excessively persistent inequality may be a long-lasting outcome of ancestors' self-selection into migration, and need not involve e.g. ethnicity-based behaviors. A testable implication is that the correlation between home country characteristics that influence self-selection, and migrants' and their descendants' outcomes should increase generation by generation. Verifying this, their ancestors' migration distance has risen to explain around half the inequality between fourth-generation immigrant groups today.

Key words: migration; selection; intergenerational mobility; ancestry; immigrant integration
JEL codes: F22, I24, J61, J62,

* I thank Jan Stuhler, Mikael Lindahl, and Inge van den Bijgaart for valuable inputs and discussions.

1 – Introduction

Over the last two centuries the United States has been the world's prime destination for international migration, attracting tens of millions from all corners of the world. Given equal opportunities, initial differences between these different migrant groups were long expected to fade away over the course of a few generations as they all, according to the popular metaphor, melted in the same pot. Yet since a few decades back, research has pointed out important limitations to America's success in providing economic equality in the melting pot. Among others, Lieberson and Waters (1988), and Borjas (1992, 1994) have shown large socioeconomic differences depending on ancestry even in the white majority population of third- or higher-generation immigrants of European origin. This limited convergence also appears to represent something more in addition to intergenerational mobility in the United States being low in general, both compared to its mobility ideal (Solon, 1992), and to most other Western countries (Corak, 2006). Specifically, Borjas (1992, 1994) has shown that mobility is even lower between ancestry groups than between individuals.

Why then does convergence between ancestry groups take so long? Straightforward and popular interpretations revolve around some form of mobility-inhibiting behavior based on ethnic group-identification. This could be e.g. through the ethnic environment outside of the family directly transmitting knowledge, ideals, role models etc. to children growing up (Borjas' "ethnic capital" hypothesis), or to ethnic discrimination (Borjas, 1992; Solon, 2014). Yet existing empirical evidence does not specifically favor any particular interpretation, and the question of the reason for the low mobility of ancestry groups remains largely open.

This paper argues that excess persistence of ancestry-group inequality need not represent a failure of ethnic relations or current policies, or even be at all related to ethnic group-identification. Instead this persistence may be a long-lasting outcome of how ancestors once self-selected into migration to America. The transmission mechanisms by which this self-selection remains important today are then the normal mechanisms of nature and nurture by which economic status is transmitted from parents to children within families. No ethnicity-based identification or behaviors are required.

The theoretical starting point for this conclusion is that economic outcomes are due to multiple underlying endowments at the individual level. These may include e.g. preferences and motivation, cognitive and non-cognitive abilities, health, school and teacher quality, and chance. Some of these endowments are more strongly inherited than others from parents to

children. The ordering is not known, but plausibly preferences and abilities are among the most strongly, and chance the least strongly inherited. The degree to which children inherit their parents' outcomes therefore depends on the sources of these outcomes. Specifically, if inheritance rates are independent, inheritance of observed outcomes will be a weighted average of the inheritance rates of the underlying endowments, with weights that correspond to the shares of outcome variance in the parent generation that are due to each factor (Lefgren, Sims, and Lindquist, 2012). There is little reason for these weights to be identical at the ancestry-group and individual levels. Hence observed outcome mobility may plausibly differ in either direction between the two levels.

To furthermore see the likely direction of the difference, we may note that most individuals, today as well as one-hundred years ago, self-select into migrating to America in order to maximize their income. As the distributions of returns and costs of migration differ between countries of origin, these self-selected samples are drawn differently from different populations (Sjaastad, 1962; Chiswick, 1978; Borjas, 1987; Grogger and Hanson, 2011). This is typically considered the primary reason for economic inequality between first-generation immigrants from different countries. Which individual traits are most important in the self-selection process is not known. Yet traits like those just mentioned as plausibly more strongly inherited than others (preferences, cognitive and non-cognitive abilities), are typically thought to be central (Chiswick, 1978; Borjas, 1987). Hence the more persistent sources of individual outcomes would make up larger fractions of outcome variance between ancestry groups than between individuals, i.e. creating higher observed outcome persistence at the ancestry-group level.

Existing empirical evidence is neither more nor less consistent with this compared with other explanations for the low mobility of ancestry groups. However the explanation proposed here implies one fairly demanding testable prediction that does not follow from explanations based on ethnic behaviors. Namely, the self-selection pattern of the immigrant generation – i.e. which groups are more positively or negatively selected – should become increasingly visible in each new generation of descendants of these immigrants. The reason for this rather counter-intuitive result is the following: Average economic outcomes of first-generation immigrant groups of different origins differ for multiple reasons. In addition to self-selection, there are e.g. differences in the average extent and quality of education in their home countries, and the transferability of this education to the US. They may also be subject to group-level shocks upon arrival, such as by unexpected economic downturns. Descendant generations gradually

converge towards a common level of economic outcomes. Yet in this process, the part of initial inequality that was due to self-selection on highly persistent endowments will, by definition, decline more slowly than other components that were due to less persistent endowments. The endowments that were important for self-selection will therefore explain a larger and larger share of outcomes in each new generation.

This prediction is of limited empirical value in itself, since the endowments on which migrants were self-selected are not directly observable. However, according to established models, the sign and extent of self-selection will depend on observable characteristics of the home country, whereof the two most commonly considered are its distance from the destination (Sjaastad, 1962; Schwartz, 1973), and its income inequality (Borjas, 1987). We thus have the testable prediction that these characteristics will explain larger and larger shares of outcomes in each new generation of descendants. Mainly for reasons of data availability and quality, this prediction will be evaluated in relation to the migration distance.

The empirical support for the prediction is quite striking. In the sample of male mostly fourth-generation immigrants observed in 2010-14, whose great-grandfathers immigrated around the turn of the 20th century, the great-grandfathers' migration distance explains a full 49% of total occupational prestige inequality between ancestry groups, whereas it explained only 14% among the great-grandfathers themselves. Strong support for the prediction is also obtained when following a larger number of origins from a more recent immigrant cohort for two generations only.

Section 2 of this paper provides the background and theoretical framework for the analysis. Section 3 describes the data and sample selections. The main empirical analysis is reported in Section 4. The analysis is extended in Section 5 by evaluation of two additional implications of the theoretical model, which however have the drawback that their confirmative results can also be explained by other, unrelated mechanisms. Section 6 concludes.

2 – Background and theory

The result that socioeconomic inequality between Americans of different ancestries is more persistent across generations than inequality between individuals was first reported by Borjas (1992). He 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 generation $t-1$:

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

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 did the outcome of the individual's own father.

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 coefficient of intergenerational persistence obtained is the sum of the two at the individual level, i.e. $\delta_1 = \gamma_1 + \gamma_2$ (Borjas, 1992). Hence the result that γ_2 is positive is equivalent to persistence being higher at the ancestry-group level than at the individual level. The latter result was reported by Borjas (1994), who estimated coefficients of persistence of log wage averages of male 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).

This high persistence is most commonly explained as resulting from a direct role of ethnicity, such as how children's learning is influenced by co-ethnics outside of the family, or discrimination (Borjas, 1992; Solon, 2014). Yet specific empirical evidence in favor of these interpretations does not exist to this date. In this paper I will suggest that excess persistence at the ancestry-group level need not be due to group-level behaviors, but that it can come about also in a setting where intergenerational transmission involves within-family processes only. The proposed mechanism is thus independent of the extent of ethnic identification, clustering, discrimination, etc., or even of the conception of ancestry groups as "ethnic". This feature may be appealing since excess persistence of ancestry-group inequality has been identified also in samples strongly dominated by third- or higher generation immigrants of different European origins (Borjas, 1992, 1994), i.e. groups which are geographically widely dispersed, and not typically thought of as much discriminated against in present-day America.

The analysis focuses on inheritance from father to son. An analysis of the intergenerational mobility of female immigrant groups is likely to be more complicated due to important

differences across countries of origin in attitudes to female education and labor force participation. The role of mothers of male descendants of male immigrants is also ignored. The mother's origin is likely to have a direct impact on the son's outcome, and hence we would expect faster convergence to native average economic outcomes for male immigrant groups with higher tendencies to partner with native women. Yet from the perspective of the male immigrants, partnering with native women is itself an integration outcome and possible mechanism of intergenerational mobility. It may therefore be inappropriate to control for in an analysis of mobility.[1] As such, the theoretical model is simplified to each family consisting of one individual only in each generation.

2.1 – Theoretical model

The central component of the theoretical model is that socioeconomic outcomes, such as schooling, income, or occupation prestige, have several different sources at the individual level. These include e.g. preferences and motivation, cognitive and non-cognitive abilities, health, school and teacher quality, and chance. For simplicity I allow the effects of these endowments on outcomes to be independent and linear. Letting e denote the endowments, and indexing them by k, we thus have

$$y_{it} = \sum_{k} e_{kit}$$

where endowments are measured in units of their contributions to the outcome. Allowing independence between endowments also in intergenerational transmission, each endowment is inherited from father to son by a factor ρ_k, i.e.

$$e_{kit} = \rho_k e_{kit-1} + \varepsilon_{kit}$$

Inheritance of the observed outcome, denoted β_i, is then a weighted average of the inheritance rates of the endowments. The weights, denoted σ, are the shares of outcome variance in the parent generation that are due to each endowment:[2]

$$y_{it} = \beta_i y_{it-1} + \upsilon_{it}$$

[1] Among the male immigrant groups observed in the present study, the extent of partnering with native women is strongly positively correlated with income.

[2] This result has previously been shown in a two-factor model by Lefgren, Sims, and Lindquist (2012), and in a multiple-factor model by Stuhler (2012). In addition to direct transmission of endowments, Lefgren, Sims, and Lindquist also included transmission of the outcome as such through a channel that was interpreted as a monetary effect. Adding such a channel does not alter any of the conclusions drawn here, as it simply increases all inheritance rates ρ_k by the same amount.

$$\beta_i = \sum_k \sigma_{kit-1}\rho_k$$

$$\sigma_{kit-1} = \frac{Var(e_{kit-1})}{Var(y_{it-1})}$$

For any grouping j of individuals, the inheritance rate of the group-level average of the observed outcome is a similarly weighted average of the inheritance rates ρ_k of individual endowments. Yet in this case the weights are the shares of each endowment in outcome variance at the group level and not the individual level:

$$y_{jt} = \beta_j y_{jt-1} + \upsilon_{jt}$$

$$\beta_j = \sum_k \sigma_{kjt-1}\rho_k$$

$$\sigma_{kjt-1} = \frac{Var(e_{kjt-1})}{Var(y_{jt-1})}$$

For most group structures that are of social and economic interest, there is little reason to expect these variance shares to be identical to those at the individual level, i.e. to expect $\beta_j = \beta_i$. On the contrary, group structures whose delineation is based on a parameter of economic interest would often be expected to have higher intra-group correlations in some individual endowments and lower in others. Group-level inequality may thus be more or less persistent than individual-level inequality, depending on whether the groups are predominantly defined by more or less persistent endowments. For certain group structures, important group-level processes may be at work in addition, further increasing or decreasing β_j. Yet while $\beta_j \neq \beta_i$ was impossible in the absence of such processes in a simpler model with only one channel of intergenerational transmission within the family, it is instead likely when we allow for multiple such channels.

The question is then which is the likely direction of the difference $\beta_j - \beta_i$ for the groups that are the focus of this study, i.e. American ancestry groups. Inequality between these groups is predominantly generated in the first generation, i.e. between immigrant groups of different origins. Descendant generations converge towards common native levels of socioeconomic outcomes, without substantial new inequality being generated along the way. What is typically considered the most important source of inequality between American first-

generation immigrant groups of different origins is that these groups are differently self-selected into migration, depending on their productive endowments, the returns to these endowments in the home country and in the US, and the monetary and non-monetary costs of moving between the two countries. Their resulting average endowment levels are therefore different depending on the characteristics of their countries of origin (Chiswick, 1978; Borjas, 1987; Grogger and Hanson, 2011).

If individuals self-select into groups depending on their endowments, group-level variance is created in the endowments on which they self-select, yet not in those on which they don't (assuming for simplicity that these are uncorrelated). Group-level variance in observable outcomes will then be entirely due to the former. If, as is believed to be the case for American immigrant groups, self-selection is not the sole but the dominant reason for group-level inequality, endowments that influence self-selection will not explain all group-level outcome variance, but the share they explain will be larger than it is for individual-level variance.

Whether ancestry-group inequality should be expected to be more or less persistent than individual-level inequality therefore depends on whether migrants' self-selection is predominantly on more or less persistent endowments. Due to their unobservable nature, which endowments are most important for this self-selection is not known, yet preferences and motivation to do well, and cognitive and non-cognitive abilities are typically believed to be central (Chiswick, 1978; Borjas, 1987). These endowments were previously mentioned as plausibly among those that are most strongly transmitted from parents to children. Hence in sum we obtain the prediction that inequality between ancestry groups is likely to be more persistent than inequality between individuals, because it is to a larger extent due to more persistent characteristics.

2.1.1 – Relation to previous literature

The point that if strongly inheritable endowments are correlated within certain groups, these groups will have lower intergenerational mobility than individuals, has formerly been made by Clark (2014). Clark lists multiple possible candidates for such, at one point (page 111) mentioning "ethnic" groups, yet does not further explore what makes different groups plausible or whether some may be more plausible than others. In relation to this, the present paper argues that immigrant/ancestry groups are particularly plausible candidates because of how they are self-selected. Clark furthermore makes his point in a more restrictive theoretical framework that can be seen as a special case of that presented here, where each outcome is

due to exactly two individual endowments, and children only inherit one of these two (the second can be interpreted as luck). Conceptually, this special case is sufficient to generate also the predictions of the present paper. Yet with multiple endowments we also see more realistically, countering some of the criticisms of Clark's argument (Chetty et al., 2014; Solon, forthcoming; Torche and Corvalan, forthcoming), that different group structures may feature low intergenerational mobility for different reasons, because they have high intra-group correlations in different endowments. We also see that certain group structures may feature *higher* mobility than individuals, which is easily illustrated e.g. by the group structure that consists of the two groups "people who won millions on lotteries" and "people who did not", i.e. where group differences are mostly due to an endowment – luck – with low intergenerational persistence.

Braun and Stuhler (forthcoming) have also used a framework similar to Clark's to argue for a non-causal interpretation of intergenerational "grandparent effects", i.e. the finding that children's outcomes tend to be positively correlated with those of grandparents or other non-nuclear family members, also after controlling for the outcomes of parents (Long and Ferrie, 2013; Lindahl et al., 2015; Braun and Stuhler, forthcoming). These "grandparent effects" and the "ethnic effects" estimated by Borjas (1992) are conceptually identical in the sense that the nuclear family belongs to a larger group, and average outcomes of this group are correlated with the outcomes of children also after controlling for those of parents. Similarly in both cases, these correlations may then be interpreted as due either to a direct social effect of the group on children's human capital accumulation, or to intra-group correlations in highly persistent individual endowments.

2.2 – Deriving a testable implication

Existing empirical evidence is not informative on whether the explanation proposed here for the low intergenerational mobility of American ancestry groups is more or less plausible than explanations that focus on direct social ethnicity-based effects. However the framework presented here also implies additional predictions that are empirically testable, and that do not follow from the ethnicity-based explanations. Yet certain of these predictions could also result from other, unrelated processes, and are therefore only explained and evaluated in the extension section of this paper. The rest of the main part of the paper derives and tests empirically one implied prediction that is difficult to obtain through other mechanisms than those proposed here.

To arrive at this prediction, we first note that in the first generation of immigrants there is socioeconomic inequality between groups of different origins for multiple reasons. First, and probably most importantly, there is self-selection (Chiswick, 1978; Borjas, 1987; Grogger and Hanson, 2011). Yet also if migrants were random samples from their home countries, their outcomes in the US would differ, because their home countries differ in factors such as the average extent and quality of education, and the transferability of this education to the US. It is also possible that groups experience different shocks upon arrival in the US, such as unexpected economic downturns. In sum, we can partition a group's outcomes in the first generation (t=1), as well as the group level averages of each of the endowments that sum to this outcome, into one component due to self-selection, indicated by a superscript s, and one "other" component, indicated by an o:

$$y_{j1} = \sum_k e_{kj1} = y^s_{j1} + y^o_{j1} = \sum_k e^s_{kj1} + \sum_k e^o_{kj1}$$

These two outcome components are inherited according to separate βj parameters, depending on their respective variance shares that are due to different endowments:

$$y_{jt} = \beta^s_j y^s_{jt-1} + \beta^o_j y^o_{jt-1}$$

where

$$\beta^s_j = \sum_k \sigma^s_{kjt-1}\rho_k$$

$$\sigma^s_{kjt-1} = \frac{Var\left(e^s_{kjt-1}\right)}{Var\left(y^s_{jt-1}\right)}$$

and correspondingly for β^o_j.

If the two sets of variance shares $\sigma^s_{kjt-1}$and σ^o_{kjt-1} are not identical, the self-selection and other components are thus differently strongly inherited. Hence, as the generations of descendants of immigrants gradually converge towards a common average outcome level, this convergence happens at different speeds for the two components. As in the present case we believe that more persistent endowments make up a particularly large share of the variance of the self-selection component, we expect this component to decline more slowly over the generations, i.e. $\beta^s_j > \beta^o_j$.

The share of total between-group variance in generation t that is due to the self-selection of the first generation is

$$\frac{\left(\beta_j^s\right)^{2(t-1)} Var\left(y_{j1}^s\right)}{\left(\beta_j^s\right)^{2(t-1)} Var\left(y_{j1}^s\right) + \left(\beta_j^o\right)^{2(t-1)} Var\left(y_{j1}^o\right)}$$

As we expect the numerator of this expression to decline at a slower rate than the denominator, the share of outcome variance that is due to self-selection will thus increase generation by generation. In other words, which migrant groups were more positively vs. more negatively self-selected into migration will be more visible among their children than among the migrants themselves, and even more visible among their grandchildren. As groups converge, the effect of self-selection on outcomes (the numerator) declines, yet the effects of other sources (the second term in the denominator) declines faster, i.e. making the effect of self-selection more visible. To the extent that the other factors are negative in character, as they probably most often are (e.g. poorer or less relevant education in the home country, negative arrival shocks), this can be interpreted as the children realizing the parents' inherent potential more fully than the parents do themselves. The children do not inherit all of the parents' potential, yet they inherit it to a larger extent than they do the obstacles that hold the parents back.

This prediction is not directly testable in itself, since the "selection" and "other" components are not observable. However, the reason why migrants from different countries are differently self-selected is that their home countries are different, and they are so to a large extent on characteristics that are observable. Hence we are in the unusual and fortunate situation that what is unobservable at the individual level can be observed by proxy at the group level, i.e. by home-country-level characteristics that influence self-selection. We thus arrive at the empirically testable prediction that the share of outcome variance between ancestry groups in generation t that can be explained by such a home-country characteristic measured at the time of ancestors' migration, i.e. R^2 from a regression of y_{jt} on the characteristic in question, should increase generation by generation. Notably, this strategy does not require that the selection proxy is orthogonal to the "other" component of outcome variance. What is required is that it is sufficiently more strongly correlated with the self-selection component.[3]

[3] An exact, and slightly involved, sufficiency condition can be derived, yet this condition is quite uninteresting since whether it holds or not is not separately testable. The strategy to test for increasing R^2 is conceptually similar to that of Lefgren, Sims, and Lindquist (2012), which in the present setting would imply investigating

This prediction has the important advantage of being difficult to arrive at through other mechanisms than that proposed here. Measurement error should primarily work in the opposite direction, towards a decreasing correlation over time, because of noise in the linking of native-born descendants of immigrants to their fathers or forefathers (see Section 3). One alternative mechanism that could make R^2 increase though is a case where migrants are self-selected only partly on their own motivation and ability to succeed in America during their own working life, and partly on their altruism towards their yet unborn (at the time of migration) children, where these two factors have high complementarity, and where parental time investment is important for children's future outcomes. In this case, individuals in the first generation do not achieve the full potential inherent in their endowments because they choose to. They work less because they spend their time investing in their children. The children thus subsequently to a larger extent realize the parents' full potential, increasing the outcome variance that can be explained by parents' self-selection. Yet if this mechanism is realistic in practice, it is arguably mostly so for female migrants. Female dominance in taking care of children is a fact in all migrant groups, and labor force participation rates and working hours are high for male migrants of all groups. Furthermore, the mechanism is only relevant as a competing explanation when applied to the mobility of the second generation. For it to explain also what happens between later generations, the altruism towards children needs to be strongly inherited from parents to children, implying that the mechanism is no longer separate from, but rather a part of, the explanation proposed in this paper.

2.2.1 – Determinants of self-selection

Two main candidates for the selection proxy emerge from the literature on migrant self-selection: the home country's income inequality (relative to that in the destination country), and its distance from the destination country, i.e. from the US. According to Borjas (1987), those who leave a country where returns to productive endowments are low will be more positively selected, as they emigrate to obtain higher returns to their high endowments, whereas for the same reason those who leave a country where returns are high will be

---

whether the coefficient in a regression of y_{jt} on y_{jt-1} increases if the latter is instrumented by the selection proxy. Technically, the only difference between the two strategies is a slight difference in the sufficiency conditions for how much more strongly correlated the instrument needs to be with the self-selection component than with the other component. In the present setting, the increasing R^2 version has the advantages of a more intuitive interpretation that is closer to the relevant aspects of the theoretical model, and that the results can be effectively illustrated graphically. Similar results to those that are reported in this paper have also been obtained using the instrumentation strategy.

negatively selected. Absent direct measures of these returns, Borjas suggests proxying them by measures of the home country's income inequality.

In the present setting, a main drawback of this candidate is data quality and coverage. Coverage is poor already in 1980 when I observe the most recent immigrant cohort, and it is nonexistent in 1930 when I observe the earliest cohort. Hence in the analysis of this paper self-selection is proxied by the migration distance. According to e.g. Sjaastad (1962), and Schwartz (1973), migrants will be more positively self-selected when they move a longer distance. The reason is that the higher monetary and non-monetary migration costs that come with a longer distance imply that higher expected returns are required for migration to be feasible, and the expected returns will be higher for those with higher levels of productive endowments. Schwartz (1973) identifies the expected positive correlation for internal migrants in the US. Yet the role of distance in self-selection has received less interest in the more recent literature on international migration. Correlations between the migration distance and outcomes of first-generation immigrants will therefore be separately reported in the empirical section of this paper.[4]

3 – Data and sample selections

The empirical analysis uses data from multiple years of censuses, American Community Surveys (ACS), and Current Population Surveys (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 give the required statistical power. They do not however provide any possibility of linking individual outcomes in one generation to the outcomes of these individuals' actual fathers. 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

[4] As is indicated in this analysis, and has been verified in additional analyses that are not reported, the migration distance also appears to be a stronger and more robust predictor of migrants' outcomes in the US than inequality measures, when studying migrants from a more recent period when coverage of the latter variables is better. Whether the predicted relations between self-selection and each of the two proxy variables are conflicting or not depends on small details of the theoretical models that are not empirically verifiable. See Chiswick (1999), and Grogger and Hanson (2011) for discussions. An additional potential proxy candidate would be the average income level of the home country (Grogger and Hanson, 2011). Yet a main drawback of this candidate is that it is obviously strongly correlated also with the "other" component of migrants' outcomes, e.g. through its strong correlation with the average education level of the home country. Empirically, its correlations with migrants' outcomes are also lower than those of the other two candidates.

the fathers of native-born individuals, with a father from country *j*, who are 25-60 years old approximately thirty years later.[5]

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 is Austria-Hungary, which ceased to exist before the sample period began, and only few respondents report their origin in Austria-Hungary rather than in any of its constituent countries. The few that do are ascribed to Austria.

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 of these 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 a substantially larger number of countries of origin compared to one or two decades earlier. Yet 1980 is still early enough to enable observation of the 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.

[5] Return migration of individuals in the first generation after being observed in the census could bias the results, yet is likely to be negligible as what is observed are stocks of immigrants most of whom have been in the US a fairly long time. Hence their probabilities of return migration after observation are low.

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.

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 these variables are lacking for the first generation, 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. I also investigate results for years of schooling and log weekly wages of generations 2-4, for which 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 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. 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.[6]

One empirical concern with self-reported ancestry is selective reporting. Individuals with multiple ancestries may identify more strongly with the ancestry that corresponds best to their level of economic success, increasing the estimated intergenerational correlations at the ancestry-group level. Alternatively, more successful individuals may be more likely to not report any non-American ancestry if their actual ancestry group is a less successful one. Importantly though, while such selective reporting could in principle explain immigrant groups' low observed intergenerational mobility as such beyond the second generation, it would not induce an increasing correlation over the generations between contemporary outcomes and the ancestors' migration distance, i.e. the prediction to be evaluated in this paper. If e.g. more successful individuals were less inclined to report belonging to less successful ancestry groups, they would determine which groups were "less successful" based on what they have been able to observe, i.e. y_{jt-1} (or some other y_{jt-x}). This would only contribute to cementing previously observed patterns, whereas an increasing correlation with the migration distance requires that the patterns change towards something they have not been able to observe, as y^s_{jt-1} (or y^s_{jt-x}) is not observable.

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 in the CPS)[7] 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

[6] In the 1990 census, 82% of the native-born sample aged 25-60 reported a non-American ancestry (and 2% reported "American Indian"). Among these, only 37% also reported a second ancestry.
[7] 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.

second-generation immigrants in the native-born sample by ancestry in 1990 is estimated to be larger than one-fourth.[8]

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 0.01–0.21, from where there is a large jump up to the 23rd lowest share at 0.33, and already the 27th is above 0.5. 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 drawn from another pattern that can be seen in Figure 1, i.e. that a very large share of pre-1930 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 estimate 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 estimate 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

[8] 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 more appropriate 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 – Descriptive overview of inequality and intergenerational persistence

Table 1 summarizes inequality between ancestry groups by cohort, generation, and outcome. A clear message conveyed by the table is that convergence between ancestry groups is slow. In fact, while we generally observe education levels and prestige scores converging (to the extent that prestige score variance is comparable when the mean changes over time), for wages we even observe slight divergence in both cohorts. A straightforward interpretation of these patterns is that levels of productive endowments have indeed converged throughout the sample periods for both cohorts, yet this convergence has not been strong enough to keep up the pace with the increasing returns to these endowments over time, which has been thoroughly documented (Autor, Katz, and Kearney, 2008).

Finally, although we observe convergence in schooling in both cohorts, substantial inequality remains in the latest cohorts also for this outcome variable. Among second-generation immigrants of the 1980 cohort, the gap between the most and the least educated groups is a

full four years, while it is two years among fourth-generation immigrants of the 1930 cohort. The corresponding wage gaps are 160% and 40% respectively.

An illustration of the high intergenerational correlations in ancestry-group inequality 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 $0.36^{1/3}=0.71$.[9]

A wider range of estimates of intergenerational persistence is reported in Table 2, with regression coefficients in column (1) and correlation coefficients in column (2). The outcome and generation pair are indicated to the left on each row. Column (3) reports the coefficients of intergenerational persistence (β) implied by the regression estimates in (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, where on the other hand the corresponding correlation coefficient is a full 0.82.

4 – Empirical analysis

I first report in some detail the relation between the self-selection proxy, i.e. the migration distance, and outcomes in the first generation of immigrants. I then proceed to evaluating the prediction that the migration distance explains increasing shares of outcomes in descendant generations. Results for the first generation are reported in Table 3. Looking first in Panel A we see in the first column that a longer migration distance is correlated with significantly higher occupational prestige scores in the 1930 cohort. Likewise the second column shows that in the 1980 cohort, an additional approximately 4,000 kilometers of migration distance is correlated with an additional year of schooling. This estimate changes little in column (3), where the average level of schooling in the home country is controlled for. It is also highly similar in column (4), where also the Gini coefficient is controlled for, as a measure of the home country's income inequality. To do this with reasonable coverage, column (4) focuses on a separate immigrant cohort that is observed in 2010.[10] Columns (5)-(7) report results

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

[10] Highly similar results are obtained if I use the income share held by the highest 20% instead of the Gini coefficient. These results are not reported.

corresponding to those of columns (2)-(4), with immigrant groups' log weekly wages as the dependent variable. These results are also highly similar, with an additional 1,000 kilometers of distance implying approximately 2% higher weekly wages.

A large fraction of the variation in migration distance is between continents. To verify that the correlations between distance and migrants' outcomes are not only due to variation between continents, I have also added dummy variables for three of the four major continents Latin America / the Caribbean, Europe, Asia, and Africa to the regression specifications that were reported in Panel A. Countries that are not part of any of these continents are excluded from these regressions. The results are reported in Panel B. All the estimated coefficients on distance are of fairly similar magnitudes to those in Panel A. The largest difference is in column (1), where including the continent dummies even makes the coefficient on distance almost twice as large.

The results reported in Table 3 also change very little if the sample is restricted to only recently arrived immigrants, hence they do not to a large extent reflect group differences in education received post-migration. These results are not reported.

4.1 – Migration distance and outcomes in later generations

An evaluation of the prediction that the migration distance explains larger shares of 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 considerably stronger in the generation of these migrants' children in 2010, where the parents' migration distance explains a full 53% of inequality between ancestry groups.[11] 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.[12]

[11] The left Panel of Figure 3 has two distinct outliers: Cambodia and Laos. These are easily explained by an important omitted variable. Both these groups are as good as entirely made up of refugees of war. Clearly, these have migrated for reasons other than high expected economic gains. Hence they should not be expected to have as high productive characteristics as their long migration distances alone would imply. If these two countries are excluded from the analysis, R^2 instead rises from 0.42 in the first generation to 0.68 in the second.

[12] Without Cambodia and Laos, this increase is from 0.25 to 0.40.

A closer inspection of Figure 3 also reveals that the residuals from the linear regression lines included in the two graphs are strongly correlated. The correlation coefficient is a full 0.80. Hence these residuals are not random noise around the regression line. Instead, as was part of the prediction derived in Section 2.2, part of the residual from the first generation remains in the second, 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.

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 (see Table 3), 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 of this sample, 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.

Finally Figure 5 reports the corresponding results for the first four generations of the 1930 cohort for the smaller sample of 22 groups that can be well enough identified also in the third and fourth generations. 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 over time – i.e. from regressing the four R^2 values on a linear time trend – is 0.017 (the non-parametric Spearman correlation has a p value of 0.027), and the p value for the difference between the R^2 values of the first and fourth generations is 0.028, each value based on 10,000 bootstrap replications. 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, i.e. that between the first and fourth generations, is 0.54.

The results reported in Figure 5 are highly similar if the three non-European ancestries are excluded, with R^2 rising from below 0.01 in the first generation to 0.39 in the fourth. As can

be seen in the figure, the European origins seem to converge to a somewhat steeper regression line than that implied in the full sample. This indicates, as is plausible, that some omitted characteristic shifts the relation between distance and outcomes downwards for the non-European compared with the European origins. The results in Figure 5 are also 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 (i.e. generations 2-4), they also look highly similar for schooling or log wage outcomes as they do for prestige scores. R^2 in the second, third, and fourth generations are 0.09, 0.35, and 0.47 for schooling, and 0.02, 0.38, and 0.41 for log wages.

We have thus obtained strong support for the prediction that the share of ancestry group inequality that is explained by the migration distance increases over the generations, and we have done so for two different immigrant cohorts that are separated by fifty years of time.

5 – Extensions

In this Section I derive and test two additional empirical implications of the theoretical framework proposed in Section 2. Compared with the prediction that was evaluated in Section 4.1, these share the drawback that they can also easily be obtained by other unrelated mechanisms, which will be explained.

5.1 – Migration distance and mobility

One implication of the theoretical model is that the migration distance should positively explain a group's absolute intergenerational mobility, i.e. it should be positively correlated with the group's average outcomes in generation t, conditional on its outcomes in t-1. As was stated in Section 2.2, we can partition group outcomes into one component due to self-selection, with a superscript s, and one due to other factors. If it was possible to accurately measure the self-selection component and run the regression

$$y_{jt} = \alpha_0 + \alpha_1 y_{jt-1} + \alpha_2 y^s_{jt-1} + \varphi_{jt}$$

the parameter α_1 would estimate the intergenerational persistence of the "other" component, whereas α_2 would estimate the difference between this rate and that of the self-selection component. Hence we would expect α_2 to be positive. Similar to what was said in relation to the prediction of an increasing R^2, this expectation remains if the selection component is

measured by a proxy variable, as long as the proxy variable is sufficiently more strongly correlated with the self-selection component than with the other component.

This prediction is thus closely related to that of the increasing R^2. Yet, in contrast to the latter, it can also easily be generated by an alternative mechanism. A positive estimate of α_2 may appear if y_{jt-1} does not give a full account of the socioeconomic status of generation t-1, i.e. in the case where if additional variables were included, they would give a more complete picture together. The self-selection proxy, i.e. the migration distance, will then most likely be correlated with these omitted variables, which may induce a positive estimate of α_2.[13] Notably though, R^2 would still decline over time.

The prediction that the migration distance is positively correlated with group outcomes in t conditional on those in t-1 is evaluated empirically in Table 4. The prediction is consistently supported. All estimated coefficients on migration distance in the table are positive and significant. They are also consistently so if regressions for the third and fourth generations of the 1930 cohort control for outcomes in t-2 or t-3 instead of t-1. These results are not reported.

5.2 – Mobility within ancestry groups

An additional implication of the theoretical model is that the persistence of inequality *within* ancestry groups should be lower from the immigrant generation to their children than between later generations. As was stated in Section 2, a key element of the proposed explanation for the high intergenerational persistence of inequality between ancestry groups is that migrants are strongly self-selected on highly persistent characteristics. A migrant group thus largely consists of individuals that are drawn from a limited range of the population distribution of these characteristics, whereas their variation in other productive characteristics may correspond more closely to that in the total population of their home country. Hence the share of within-group variance in the first generation that is due to more persistent characteristics will be particularly low, and hence observed within-group mobility will be high. In later generations, within-group variance also in the more persistent characteristics will approach the total population values, and within-group mobility will fall.

[13] Conceptually (see Warren and Hauser, 1997; Braun and Stuhler, forthcoming), if a positive estimate of α_2 is due to this mechanism, the estimate should decline if multiple indicators of socioeconomic status in t-1 are added to the regression. Yet exploring this is not feasible with the small number of observations available in the present case.

We thus obtain the testable prediction that intergenerational mobility within ancestry groups should be higher from the first to the second generation of immigrants, than between later native-born generations. Notably though, also other mechanisms could generate this pattern. It is possible to imagine e.g. that foreign-born parents have a weaker role model impact on their native-born children, and less relevant human capital to transmit, compared with native-born parents.

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 this analysis. To enable an empirical evaluation, 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 (similar to Borjas, 1992), 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. This sample consists of all men aged 25 or 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, I regress outcomes of native-born men on the outcomes of their fathers, adding ancestry dummies to make the estimated coefficients of persistence refer to persistence within ancestry groups. I do this separately for those with native-born and those with foreign-born fathers and compare the coefficients obtained. The results are reported in Table 5. 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.[14]

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

[14] 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.

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 own and their father's occupations 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 at least for the part of the analysis that was based on the larger sample this result is highly significant.

6 – Conclusion

Inequality between Americans of different ancestries is highly persistent over multiple generations. This was shown in several studies in the 1980s and 1990s, and has been confirmed with more recent data in the present study. In the sample of mostly fourth-generation immigrants observed in 2010-14, the standard deviation in years of education across origins is 0.5, that in log wages is 0.08, and the inequality pattern is highly correlated with those in previous generations.

While previous explanations for these patterns predominantly focus on ethnicity-based behaviors, the present paper has argued that excess intergenerational persistence at the ancestry-group level is likely to occur also in the absence of group-based identification or behavior. The reason is that the first generation of immigrants is strongly self-selected on productive endowments that are more persistent than observable outcomes across generations. Furthermore, this self-selection looks different for different groups, implying that highly persistent characteristics make up larger shares of inequality across ancestry groups compared with across individuals. This alone is sufficient to explain excess persistence at the ancestry-group level. The explanation receives empirical support from the observation that a proxy variable for migrants' self-selection– i.e. the distance from the home country to the US – explains a higher share of group-level outcomes in later compared with earlier generations. This pattern is predicted by the model, yet difficult to come about through other mechanisms.

A crucial message contained in this result is that the excess intergenerational persistence of inequality between American ancestry groups is not necessarily an outcome of current mobility-inhibiting behavior holding individuals back from realizing their full potential because of their ethnicity. It is thus not necessarily a sign that American society could become

better off in this respect by changing current behaviors or policies. Yet more fundamentally, the result also implies that it is not even clear that low ancestry-group mobility should be seen as undesirable. When thinking about low mobility, one often tends to focus predominantly on the bottom half, i.e. that those who achieve below average in one generation continue to do so in the next. However, the same is true for those who achieve above average outcomes, and in America in particular not few of these, individuals as well as country groups, appear to be strongly positively selected. For example, among the 96 newly-arrived (last ten years) male immigrant groups that enable a comparison in 1980, one-third had on average more than ten years more education than the corresponding adult male averages of their home countries. Without much doubt, America provides the world's highest returns to exceptional abilities and willingness to work hard for economic success, and has for a long time succeeded better than any other country in attracting such individuals from the rest of the world. Part of the conclusion from the present analysis is that these individuals' qualities have remained or will remain to a large extent in several generations of their descendants, to America's long-term benefit.

References

Autor, David, Lawrence Katz, and Melissa Kearney (2008), "Trends in U.S. wage inequality: Revising the revisionists", *Review of Economics and Statistics*, 90: 300-323

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*

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

Corak, Miles (2006), "Do poor children become poor adults? Lessons from a cross-country comparison in generational earnings mobility", in John Creedy and Guyonne Kalb (eds): *Dynamics of Inequality and Poverty*, https://doi.org/10.1016/S1049-2585(06)13006-9

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

Lefgren, Lars, David Sims, and Matthew Lindquist (2012), "Rich dad, smart dad: Decomposing the intergenerational transmission of income", *Journal of Political Economy*, 120: 268-303

Lieberson, Stanley, and Mary Waters (1988), *From many strands: Ethnic and racial groups in contemporary America*, Russel Sage Foundation

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

Long, Jason, and Joseph Ferrie (2013), "Intergenerational occupational mobility in Great Britain and the United States since 1850", *American Economic Review*, 103: 1109-1137

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 (1992), "Intergenerational income mobility in the United States", *American Economic Review*, 82: 393-408

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

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

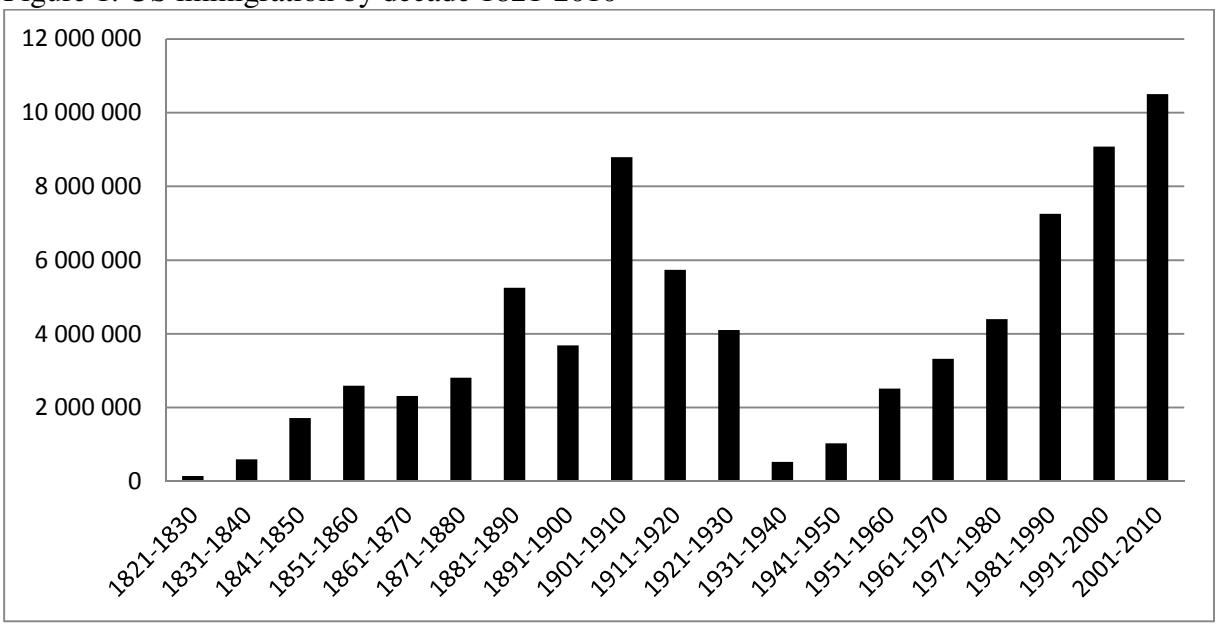
Stuhler, Jan (2012), *Mobility across multiple generations: the iterated regression fallacy*, IZA Discussion Paper No. 7072

Torche, Florencia, and Alejandro Corvalan (forthcoming), "Estimating intergenerational mobility with grouped data: a critique of Clark's The son also rises", *Sociological Methods and Research*

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

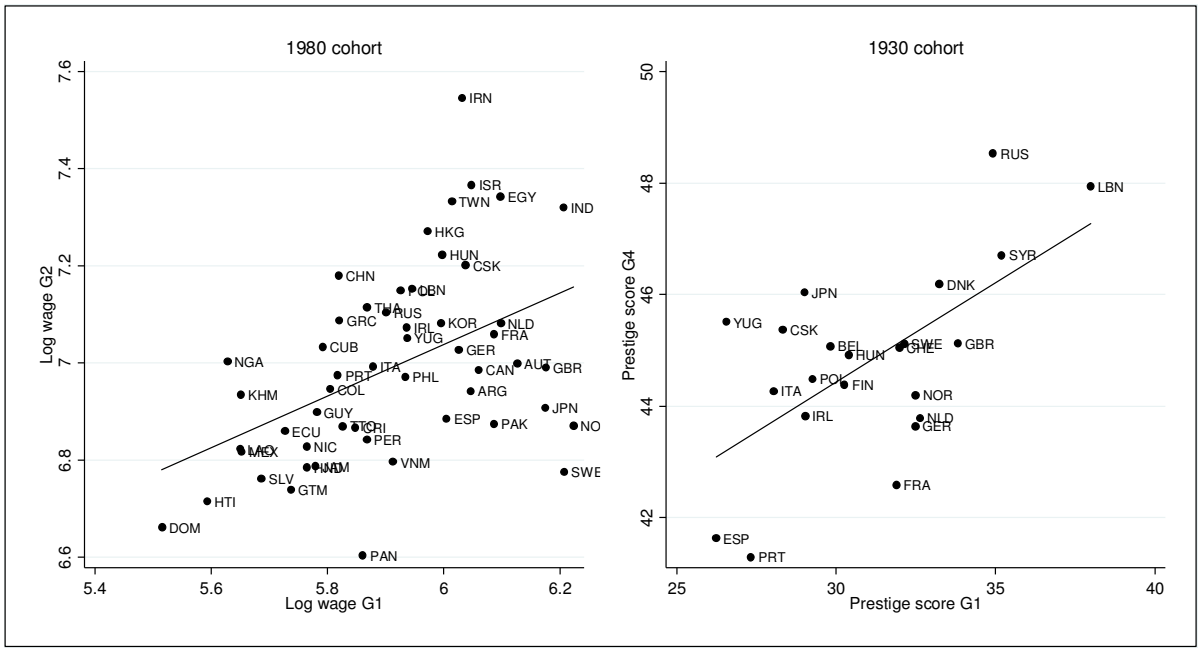
Figures

Figure 1. US immigration by decade 1821-2010



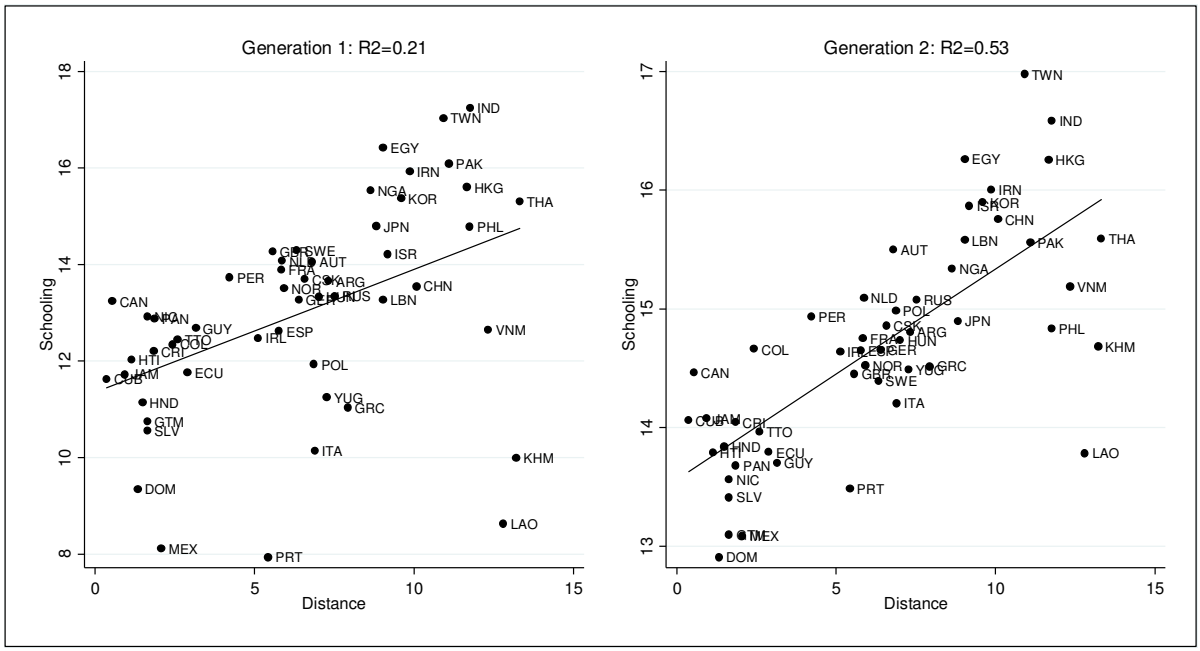
Notes: Immigration is measured in absolute numbers by decade. Data source: 2014 Yearbook of Immigration Statistics, US Department of Homeland Security

Figure 2. Intergenerational persistence of ancestry groups' socioeconomic outcomes



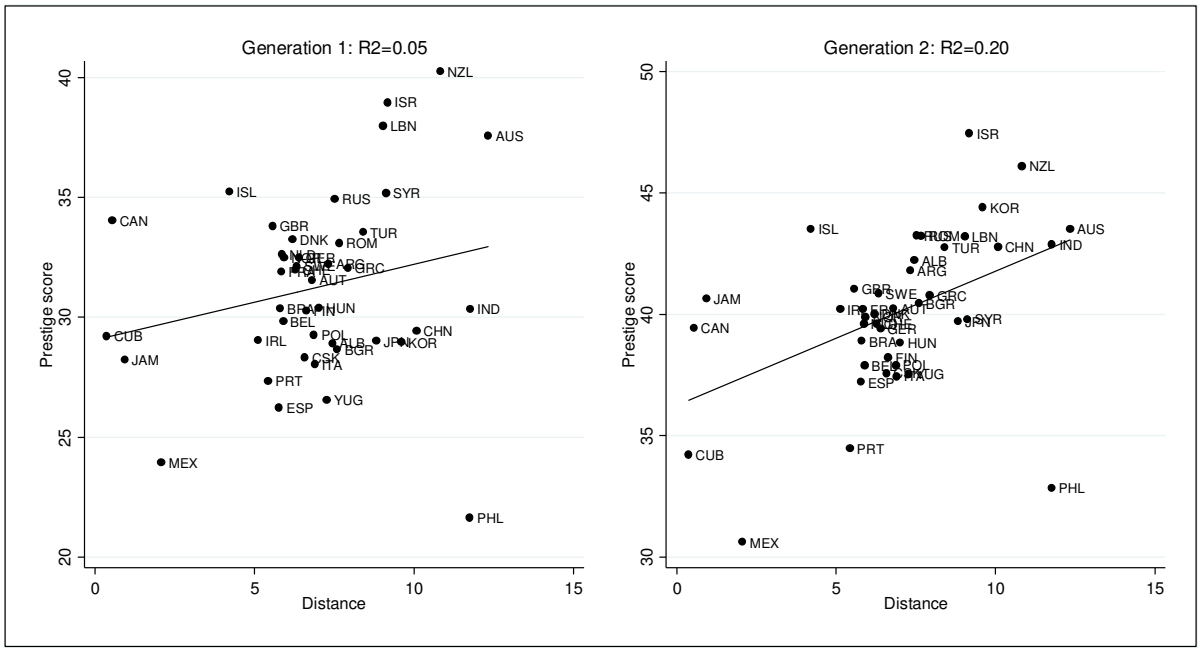
Notes: The graph to the left (n=52) shows average log wages of immigrants in 1980 (G1), and of their children in 2010 (G2) by country of origin. The graph to the right (n=22) shows average occupational prestige scores of immigrants in 1930 (G1), and of their great-grandchildren in 2012 (G4) by country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

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



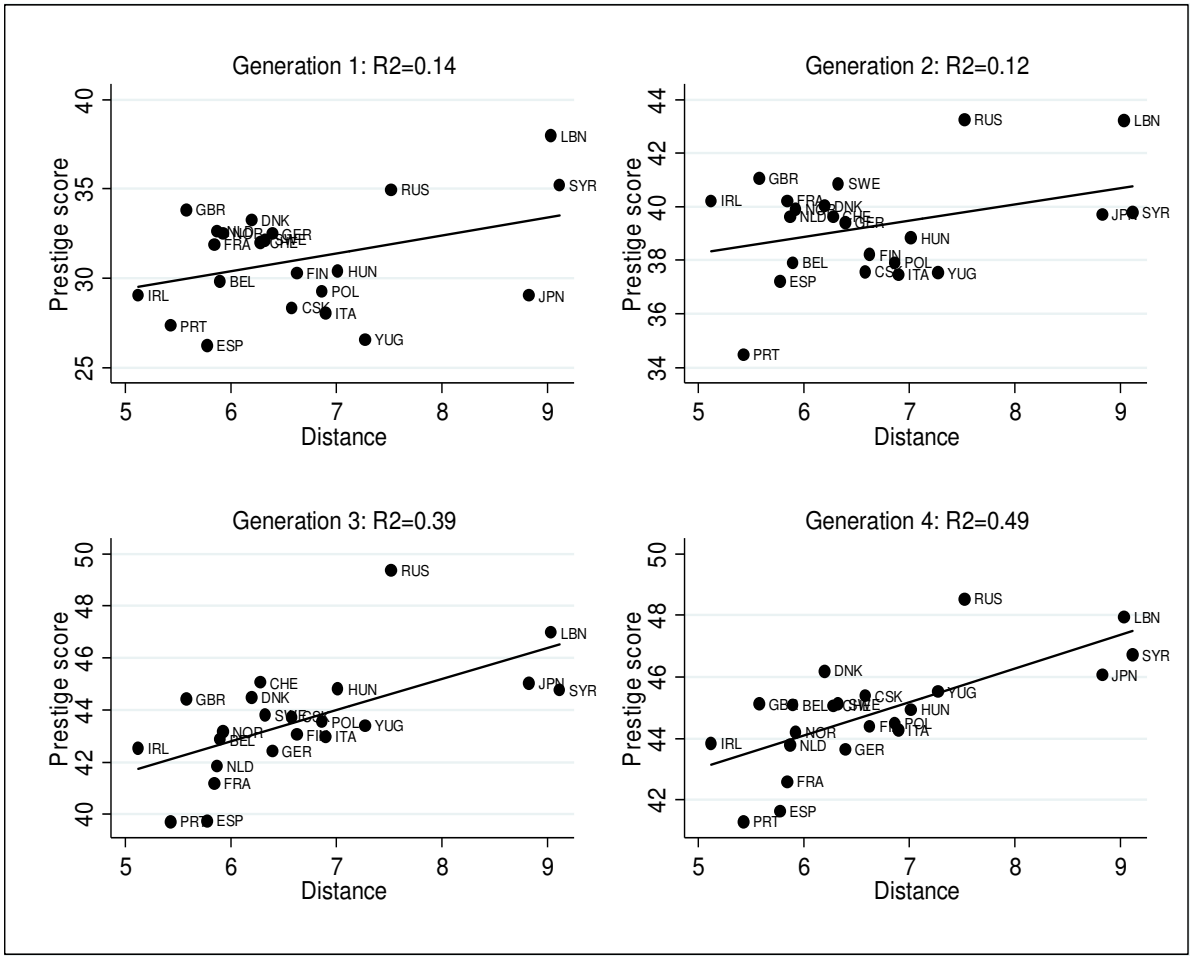
Notes: N=52. Each observation is a migrant group from one country of origin in the US. Distance is measured in thousands of kilometers.

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. Inequality descriptives by generation

	Mean	Standard dev.	Min	Max	Obs
1980 Cohort					
--- G1 Schooling	12.97	2.15	7.94	17.25	52
--- G2 Schooling	14.69	0.93	12.91	16.98	52
--- G1 Log wage	5.91	0.174	5.52	6.22	52
--- G2 Log wage	6.99	0.197	6.60	7.55	52
1930 Cohort					
--- G1 Prestige score N=42	31.22	3.78	21.66	40.24	42
--- G2 Prestige score N=42	40.07	3.31	30.62	47.47	42
--- G1 Prestige score N=22	31.05	3.01	26.24	37.98	22
--- G2 Prestige score N=22	39.27	1.98	34.47	43.25	22
--- G3 Prestige score	43.58	2.14	39.71	49.36	22
--- G4 Prestige score	44.80	1.74	41.29	48.53	22
--- G2 Schooling	11.16	0.71	9.22	12.44	22
--- G3 Schooling	14.04	0.66	12.67	15.68	22
--- G4 Schooling	14.20	0.48	13.21	15.20	22
--- G2 Log wage	4.81	0.062	4.67	4.92	22
--- G3 Log wage	6.46	0.079	6.29	6.68	22
--- G4 Log wage	6.91	0.080	6.73	7.09	22

Note: The outcome, cohort, and generation are indicated to the left on each row.

Table 2. 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**	0.821	0.355	52
	(0.035)			
--- G1-G2 Log wage	0.531**	0.469	0.531	52
	(0.147)			
1930 Cohort				
--- G1-G2 Prestige score	0.671**	0.767	0.671	42
	(0.087)			
--- G1-G3 Prestige score	0.446**	0.627	0.668	22
	(0.128)			
--- G1-G4 Prestige score	0.357**	0.617	0.710	22
	(0.109)			
--- G2-G3 Prestige score	0.794**	0.734	0.794	22
	(0.163)			
--- G2-G4 Prestige score	0.606**	0.688	0.778	22
	(0.124)			
--- G3-G4 Prestige score	0.772**	0.949	0.772	22
	(0.062)			

Notes: Each row represents the estimated coefficient of intergenerational persistence using the outcome, cohort, 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 3. Migration distance and outcomes in the first generation

	Prestige scores	Schooling			Log weekly wages		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	1930	1980	1980	2010	1980	1980	2010
				Panel A – Without continent dummies			
Distance	0.326*	0.271**	0.276**	0.259**	0.0176**	0.0198**	0.0266**
	(0.160)	(0.044)	(0.047)	(0.0451)	(0.0042)	(0.0041)	(0.0075)
Schooling			0.0691	0.277**		0.0277**	0.0707**
			(0.0628)	(0.065)		(0.0049)	(0.0101)
Gini				-0.0328			-0.0050
				(0.0179)			(0.0041)
R2	0.070	0.278	0.268	0.541	0.145	0.307	0.575
N	82	107	96	67	107	96	67
				Panel B – With continent dummies			
Distance	0.574*	0.243**	0.243**	0.209**	0.0251**	0.0256**	0.0162*
	(0.270)	(0.072)	(0.076)	(0.072)	(0.0094)	(0.0087)	(0.0079)
Schooling			0.319**	0.417**		0.0283**	0.0720**
			(0.103)	(0.112)		(0.0068)	(0.0106)
Gini				-0.0524*			-0.0003
				(0.0256)			(0.0043)
R2	0.101	0.360	0.416	0.633	0.296	0.379	0.677
N	77	99	91	64	99	91	64

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 column head. *Distance* is the distance from the home country to the US, measured in thousands of kilometers. *Schooling* is average number of years of schooling in the home country. *Gini* is the Gini coefficient in the home country. Robust standard errors in parentheses. * = p<0.05, **=p<0.01.

Table 4. Migration distance and mobility

	(1)	(2)	(3)	(4)	(5)
	1980 cohort G2	1980 cohort G2	1930 cohort G2	1930 cohort G3	1930 cohort G4
	Schooling	Log wage	Prestige score	Prestige score	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 5. Intergenerational within-group inheritance from fathers to 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.