DISCUSSION PAPER SERIES

No. 6316

INTERGENERATIONAL MOBILITY AND THE INFORMATIVE CONTENT OF SURNAMES

Maia Güell, José V Rodríguez Mora
and Chris Telmer

*INTERNATIONAL MACROECONOMICS
and LABOUR ECONOMICS*

Centre for Economic Policy Research

www.cepr.org

INTERGENERATIONAL MOBILITY AND THE INFORMATIVE CONTENT OF SURNAMES

Maia Güell, University of Edinburgh, CEP (LSE), IZA and CEPR
José V Rodríguez Mora, University of Edinburgh and CEPR
Chris Telmer, Carnegie Mellon University

CEPR Discussion Paper No. 6316

June 2007
Revised February 2012

ABSTRACT

Intergenerational Mobility and the Informative Content of Surnames*

We propose a new methodology for measuring intergenerational mobility in economic wellbeing. Our method is based on the joint distribution of surnames and economic outcomes. It circumvents the need for intergenerational panel data, a long-standing stumbling block for understanding mobility. A single cross-sectional dataset is sufficient. Our main idea is simple. If 'inheritance' is important for economic outcomes, then rare surnames should predict economic outcomes in the cross-section. This is because rare surnames are indicative of familial linkages. Of course, if the number of rare surnames is small, this won't work. But rare surnames are abundant in the highly-skewed nature of surname distributions from most Western societies. We develop a model that articulates this idea and shows that the more important is inheritance, the more informative will be surnames. This result is robust to a variety of different assumptions about fertility and mating. We apply our method using the 2001 census from Catalonia, a large region of Spain. We use educational attainment as a proxy for overall economic well-being. Our main finding is that mobility has decreased among the different generations of the 20th century. A complementary analysis based on sibling correlations confirms our results and provides a robustness check on our method. Our model and our data allow us to examine one possible explanation for the observed decrease in mobility. We find that the degree of assortative mating has increased over time. Overall, we argue that our method has promise because it can tap the vast mines of census data that are available in a heretofore unexploited manner.

Maia Güell
School of Economics
University of Edinburgh
31 Buccleeuch Place
Edinburgh
EH8 9JT

Email: maia.guell@upf.edu

For further Discussion Papers by this author see:
www.cepr.org/pubs/new-dps/dplist.asp?authorid=138442

José V Rodríguez Mora
School of Economics
University of Edinburgh
31 Buccleeuch Place
Edinburgh
EH8 9JT

Email: sevimora@gmail.com

For further Discussion Papers by this author see:
www.cepr.org/pubs/new-dps/dplist.asp?authorid=139713

Chris Telmer
Tepper School of Business
Carnegie Mellon University
249 Posner Hall
Pittsburgh, PA 15213
USA

Email: chris.telmer@cmu.edu

For further Discussion Papers by this author see:
www.cepr.org/pubs/new-dps/dplist.asp?authorid=148609

Submitted 15 May 2007
Revised 1 February 2012

1   Introduction

How mobile are income and wealth across generations? Or, in plainer language, how important is the economic well-being of parents for determining the economic well-being of their children? The answer is important for many fields of economics, ranging from public finance to education policy to savings and portfolio choice.

Ideally, we'd like to understand intergenerational mobility along (at least) three dimensions. First, at a point in time. How are the opportunities faced by today's children affected by the economic and social status of their parents? Second, across time. Has this relationship changed with respect to previous generations? Third, across countries. Is mobility higher in the U.S. than in Europe?

Unfortunately, we don't know much about the answers to these questions, in particular the second and third. The main reason is data limitations. To address the questions directly, panel data is needed, linking the economic status of adults to that of their parents for multiple generations and countries. The existence of such data is rare. That which does exist (i) has been strongly criticized in terms of various biases, (ii) is not comparable across countries, and (iii) cannot address the question of how mobility has *changed* over time (*c.f.* Solon (1992, 2002)). These last two points deserve emphasis. The complexity of intergenerational mobility makes its measurement at a point in time very difficult to do and interpret. Comparing measures of mobility across time or populations would be far less problematic... if we had the data.

This paper attempts to make headway by introducing a new source of data: *surnames* and how they vary with measures of economic well-being. We show that the implicit intergenerational links that are inherent in surnames provide a useful stand-in for the explicit intergenerational links that would exist in multi-generational panel data. We do so using both theory and data. Our theory shows how surname data can allow one to estimate both the level and the change in intergenerational mobility *even in the absence of explicit links between children and their parents.* Our empirical work implements this idea. We use novel, census-based data from Catalonia — a large region of Spain — to obtain measures of mobility both at one point in time and across generations. The former concord well with previous studies. The latter, more strikingly, provide an estimate of how mobility has *changed* over time. We find that, among the different generations of the 20th century, it has *decreased*. That is, the economic status of parents seems to have

become *more* closely related to that of their children. Why? Our methodology offers one potential explanation. Assortative mating, the tendency for people with similar economic status to mate with one another, has increased over time. We use our surname data to establish this and our model to demonstrate how it is one potential source of a decrease in intergenerational mobility.

Our main idea is straightforward. Intergenerational mobility is all about how children inherit aspects of economic well-being from their parents. In data, we can observe *outcomes* on economic well-being, but, unless the data contain explicit knowledge of who is the parent of who, we are (apparently) unable to determine the degree to which the economic well-being has been inherited. However, suppose that the data also allow us to observe surnames. Surnames are almost always inherited from one's father. Thus, they can serve as *markers*. They are intrinsically irrelevant for the determination of economic well-being, but they get passed from one generation to the next, alongside other characteristics that *do* matter. The more important are these characteristics in determining outcomes, the more inheritable are the outcomes, and, therefore, the more information the surnames will contain on the values of outcomes. In this way, surnames can be used to measure the importance of inheritance and thus identify the degree of mobility. The following example articulates this mechanism in more detail.

Consider a society comprised of two distinct groups of people: rich and poor. In each group there are males and females, but, because surnames are (typically) passed along the male lineage, we will ignore the females for now (mating is discussed later). Suppose that the males within each group all share the same surname: Richmanson for the rich and Poormanson for the poor. This means that if we partition society either by surname or by economic status, we get the same thing. We would say that, among this initial generation, surnames are 'perfectly informative.' If you know a man's surname, you know his economic status. Now, how informative will surnames be among the subsequent generation? The answer depends on the degree of inheritance. Consider two extremes. First, if inheritance is 'perfect' — meaning that there is no mobility whatsoever — then the economic status of all sons would be identical to that of their fathers. Surnames, being passed from father to son in exactly the same manner as economic status, would remain perfectly informative. Second, if inheritance is irrelevant, so that the sons of both the Richmansons and Poormansons are equally likely to be rich or poor, then surnames would become perfectly *un*informative among the next generation. Thus, the informativeness of surnames depends on the degree of intergenerational mobility. This is the essence of the mechanism with which we are able

to infer the degree of mobility.

Reality, of course, lies between these two extremes, and surnames carry some information. But, there is a tendency for surnames to become non-informative as time goes on. To understand this, suppose that economic status follows a stationary process with some degree of persistence, and that this process is the same for both the Richmansons and the Poormansons. In this case, surnames will remain informative among the sons of the initial generation, but the informativeness will not be perfect. It will become less perfect with each subsequent generation. After enough generations the cross-sectional distribution of status among both elements of the surname partition will be the same and the initial informativeness will have vanished. This is what makes our methodology somewhat less than obvious. For surnames to contain information about intergenerational mobility there must be something else going on that inhibits this convergence to a common stationary distribution.

The additional ingredient in our study is a *birth-death process* for surnames. Surnames die when the last male holder of a particular name bears no male children. They are born when someone mutates their name, or when a new name enters the population via immigration. This generates a *skewed* surname distribution, with a small number of names each being held by a large number of people and a large number of names each being held by a small number of people. The surname distribution in most Western societies takes exactly this form.

Herein lies the key to our method: skewness in the surname distribution. We can't learn anything from the name Smith. The cross-sectional distribution among the Smiths is similar to that among society as a whole (in particular if one controls for ethnicity, as we do). We can, however, learn something from the multitude of *rare* surnames. This is because they are likely to form a partition of the population that is correlated with familial linkages. Suppose, for example, that a rich person chooses a new, unique surname and that this person's male descendants maintain this name. Then, if mobility is low, this surname will be informative for some number of subsequent generations. It will be shared by a bunch of rich people. If mobility is high it will not be as informative. The same logic holds for a poor person who changes their name, or for an immigrant with a distinct name. Over time these people's surnames will either die or will multiply. If they multiply, then, as discussed above, they are likely to become less informative. But, overall, a stationary birth/death process will generate a stationary *commonality-ordered* surname frequency distribution. There will always be some rare names. These are the names from which we can extract information on intergenerational mobility.

An important issue for us is ethnicity. Early-generation immigrants, for example, tend to have distinctive surnames and relatively low economic status. Surnames in a society with lots of recent immigration, therefore, contain information on both familial linkages *and* ethnicity. The latter is likely to make surnames economically informative. We seek to isolate the former — the importance of familial linkages — and therefore we need to control for ethnicity in some way. Fortunately, in our data, this turns out to be relatively easy. In the region of Spain that we study, Catalonia, the vast majority of the immigration has either been very recent, or has originated from other regions of Spain. We control for the recent immigrants by considering only Spanish born, Spanish citizens. We control for the effect of Spanish regional origin — which is well known to be highly relevant — by using telephone directories to develop an index of the 'ethnicity' (from the perspective of Catalonia) of each surname in the population.

Our method is not without its limitations, but they are counterbalanced by some important strengths. One is that it is quite data-friendly. We are able to measure mobility from a single, cross-sectional census. We do not require any explicit links between parents and children. As a result, the censuses that are periodically compiled by many governments contain most of the information that we require. Moreover, a great deal of confidentiality and anonymity can be maintained while still allowing access to the necessary information. Surnames can be encoded without negating what we do.

Our method can be applied to any country that follows the Western naming convention. Spain, the source of our data, uses a variant of the Western convention that is *identical* to its Anglo-Saxon counterpart, but with the additional ingredient of the maternal surname, which survives for one generation. We show that knowledge of the maternal name can be exploited in three ways: (i) to control for ethnicity, (ii) to determine the degree of assortative mating among the parents of an individual, and (iii) to partition our data into sets of siblings. The latter affords us a powerful robustness check on our methodology while at the same time providing a useful link to a large literature that attempts to measure mobility based on correlations between siblings' economic outcomes.

The remainder of our paper is organized as follows. Section 2 provides an overview of the existing literature and discusses the data limitations that are the primary motivation for our new initiative and data source. Section 3 defines our measure of the 'Informational Content of Surnames' (ICS). Section 4 develops and analyzes a model determining the joint distribution of surnames

4

and incomes. Section 5 extends our baseline model in several dimensions that are important for interpreting our subsequent empirical results. Section 6 describes our data, Section 7 reports results on mobility in Catalonia in the year 2001, and Section 8 uses these results to calibrate our model. Section 9 reports results on how mobility has changed over time. Section 10 reports results on assortative mating and argues that it offers a compelling interpretation of our results on the evolution of mobility. Section 11 affirms the robustness of our methodology by constructing an alternative dataset based on siblings. Section 12 concludes.

2   Existing Literature

The handbook chapters by Solon (1999) and Black and Devereux (2011) provide comprehensive surveys of research on intergenerational mobility. Here, we limit our discussion to issues that are specifically related to empirical measurement.

Solon (1992) represents a watershed in the empirical study of intergenerational mobility. Prior to Solon's article, estimates of the correlation between parental and child income in the U.S. were relatively low, indicating high mobility. Behrmana and Taubman (1985, 1990), Becker (1967), and Becker and Tomes (1979, 1986), for instance, found estimates in the neighborhood of 0.20. Solon (1992) pointed out some problems with these estimates. He argued that they probably overstate mobility because (i) current income is a noisy proxy for lifetime income, (ii) children's income tends to be measured at the onset of their working lives which, especially for the highly educated, provides a poor approximation of lifetime income (*c.f.*, Haider and Solon (2006), Hertz (2007)), (iii) heterogeneous attrition rates and the short time dimensions of the panel datasets generate additional biases, making it very difficult to assess the dynamics in mobility. Solon (1992) addressed a number of these issues and argued for a mobility estimate in the neighborhood of 0.40. A number of subsequent studies have reached similar conclusions based on different data. A value of 0.40 seems somewhat of a consensus, at least based on U.S. data for the latter third of the 20th century.

For other countries, a number of studies exist.[1] Solon (2002), however, argues that differing

---

[1]Björklund and Jäntti (1997), Osterberg (2000), Osterbacka (2001) and Björklund et al. (2002) make use of the extensive data that is available for the Nordic countries. Dearden, Machin, and Reed (1997), Wiegand (1997), Couch and Dunn (1997) and Checchi, Ichino, and Rustichini (1999) study the U.K., Germany and Italy. Comi (2003) uses the European Community Household Panel to obtain estimates for 12 EU countries. Data is also available for Australia (Leigh (2007)), South Africa (Hertz (2001)), Brazil (Dunn (2007) and Ferreira and Veloso (2006)),

levels of the afore-mentioned biases, in addition to other sources of heterogeneity, makes cross-country comparability very problematic. Similarly, little is known about the time evolution of mobility. Lee and Solon (2006) and Hertz (2007), for example, attribute a fairly divergent body of existing results to small-sample bias in addition to the aforementioned age bias and sample attrition problems (*c.f.*, Mayer and Lopoo (2005) and Fertig (2004)). Taking this into account leaves the authors inconclusive about trends in intergenerational mobility. One exception is Blanden, Goodman, Gregg, and Machin (2004), who argue for a decrease in mobility in the U.K. between two cohorts of people born in 1958 and 1970, respectively.

Several alternative approaches to measuring mobility have been suggested in an attempt to overcome the problems associated with panel-data availability. Yet these approaches still require data with explicit family links, something which is not always available. Solon, Corcoran, Roger, and Deborah (1991) and others have studied siblings and Page and Solon (2003), Dahan and Gaviria (2001) and Levine and Mazumder (2007) have studied neighbors. Duncan, Featherman, and Duncan (1972) measure the mobility in "social prestige" associated with the professions of parents and children. Aaronson and Mazumder (2008) use large samples from the U.S. decennial Censuses and approximate parental income based on information that is available at the child-level (*e.g.*, the state of birth). Interestingly, they argue that mobility in the U.S. increased from 1950 through 1980 but has declined sharply since. Our results for Catalonia are not dissimilar.

Turning from the literature on mobility to the literature on *names*, many studies exist. Fryer and Levitt (2004), Levitt and Dubner (2005), Bertrand and Mullainathan (2004) and others study the distribution of *first names* to understand phenomenon ranging from racial discrimination to economic status. These studies exploit the *endogeneity* of first names: parents *choose* them and this choice can be related to parental characteristics. Our study, in sharp contrast, makes use of the fact that surnames are surely much more *exogenous* in nature. Wealthy people, for instance, may choose aristocratic first names for their children, but they are much more likely to simply pass along the surname that has been passed to them by previous generations. The distribution of first names follows complex social rules. The distribution of surnames, in contrast, follows simpler rules that are akin to well-known genetic laws of motion.

The use of surnames in science has a rich history, dating back to the times of Charles Darwin. Sir Francis Galton — cousin of Darwin — studied how surnames disappear in an attempt to better

Singapore (Ng (2007)), Malaysia (Lillard and Kilburn (1995)) and for several other countries (Grawe (2004)).

understand the extinction of aristocratic families. This resulted in the stochastic process of Galton-Waltson (1874). George Darwin (son of Charles) was the first one to use surnames (specifically: marital isonymy, equal surnames) in order to determine the frequency of cousin marriages in England.[2] Since then, surnames have been used for determining the degree of inbreeding inside populations, for determining population movements and determining population homogeneity (see Lasker (1985)). Spanish surnames (what we use here) have played a particularly important role, owing to the fact that they contain maternal information and that the Spanish surname mutation rate has been relatively low.

A relevant reference in biology on the mathematics of surname distribution is Manrubia and Zanette (2002). These authors consider a model of surname generation with exponential population growth. As in our model, newborn agents receive a new surname (a "mutation") with a fixed probability. With the complementary probability they are randomly assigned an existing surname from the existing set of names. In the latter case, the likelihood of receiving a particular, existing surname is proportional to that name's frequency in the existing population. In their baseline model the cross-sectional distribution of surnames (the frequency associated with which the nth most common surname) follows Zipf's law: the frequency is inversely proportional to the surname's rank. This feature of their model is consistent with data. What's inconsistent, however, is the model's time-series behavior. The number of surnames grows exponentially, at a rate determined by the mutation rate and the population growth rate. In (some) observed data the number of surnames seems to decrease with time. Motivated by this, Manrubia and Zanette (2002) add mortality risk to their model and show that under certain parametrizations the model is consistent with a shrinking set of surnames. Our model is distinct in that – by virtue of the fact that we rely on computational simulations – we keep track over time of lineages. Doing so is critical for us since we ultimately care about the joint distribution of economic characteristics and surnames, not just the marginal distribution of the latter which is the focal point of the study by Manrubia and Zanette (2002). On the other hand, our current approach does not allow for population growth, something which we leave for future work.

The use of surname data in economics has been relatively rare. Some recent exceptions include the following. Bagüés (2005) uses very long and unusual surnames in order to determine family

---

[2]Darwin (1875), as cited in Lasker (1985). Darwin was worried about the possible nocive effects of consanguinity between parents, as his father and mother were first cousins.

relationships (and the possibility of corruption) in the grades obtained in public examinations in Spain. Collado, Ortuño-Ortín, and Romeo (2006) is an attempt to distinguish the extent to which consumption behavior is inherited versus environmental. They use the distance in the distribution of surnames between provinces and do not use microdata, but only aggregated distributions. Closer to our work are Angelucci, De Giorgi, Rangel, and Rasul (2010) and Long and Ferrie (2011), who use surnames in microdata in order to identify family links in Mexico, and the US and Britain, respectively. Our objectives and methodology, however, are quite different. They use surnames as family links explicitly and intensively, (*i.e.*, to determine familial links in a small sample) while we use them implicitly and extensively for the whole population. Clark (2010) uses the distribution of surnames as a measure of long run social mobility in England. Finally, Olivetti and Paserman (2011) propose a new method that links two US censuses using *first names*. As mentioned earlier, this paper also exploits the fact that first names are endogenous. And, similar the above surname papers, the underlying idea is also to provide some explicit family links. Here these links are done in a in a more aggregate fashion which allows them to use bigger samples.

To our knowledge ours is the first paper to use surnames to study intergenerational mobility, to use the surname information from census data, and to build a theoretical framework to map the characteristics of the surname distribution into economically-interpretable units.

3   The Informational Content of Surnames

The population consists of N individuals. Each individual is associated with one surname, s, which is an element of the finite set of all possible surnames, Ω. A *census* is list with one entry per individual in the population. The ith entry records individual i's surname, a measure of their economic well-being, e_{is}, and a vector of additional characteristics, X_{is}, such as age, gender, ethnicity, place of birth, etc. We model economic inheritance as being described by

$$e_{is} = \gamma' X_{is} + y_{is} \tag{1}$$

$$y_{is} = \rho\, y_{ip} + \varepsilon_{is} \tag{2}$$

where γ is a vector of parameters, ε_{is} is an *iid* shock with variance V_ε and y_{is} is a set of unobservable traits that are passed from parents, who have traits y_{ip}, to their children. The parameter $0 \leq \rho < 1$

measures the importance of economic inheritance.

We define the *informational content of surnames* (ICS) as the difference in the (adjusted) R^2 between two regressions. The first, with R^2 denoted R_L^2, estimates equation (1) for the *average* individual with surname s

$$e_{is} = \gamma' X_{is} + b'D + \text{residual} \; , \tag{3}$$

where D is an S-vector of surname-dummy variables with $D_s = 1$ if individual i has surname s and $D_s = 0$ otherwise. Our methodology is based on the idea that as surname s becomes more infrequent it becomes more likely that this average gets taken across individuals with familial linkages, thereby providing information about economic inheritance.

The second regression mixes up the surnames so that they cannot be informative. It is

$$e_{is} \;=\; \gamma' X_{is} + b'F + \text{residual} \; , \tag{4}$$

where F is an S-vector of 'fake' dummy variables that randomly assign surnames to individuals in a manner that maintains the marginal distribution of surnames. The R^2 from this regression is denoted R_F^2. The ICS is defined as

$$\text{ICS} \equiv R_L^2 - R_F^2 \; . \tag{5}$$

The ICS is a moment of the joint distribution of surnames and economic well-being that measures the *incremental* informational content of surnames. In our model it will turn out to be monotonically increasing in the economic inheritance parameter, ρ.

The basic idea behind the ICS measure is this. Surnames define a *partition* of the population. If the surname partition is informative about familial linkages — if some individuals with the same surname come from the same family — then it can be used to measure the importance of economic inheritance. The fake-surname partition *is constructed* to have zero information about familial linkages. By comparing the relative informativeness of the two partitions, therefore, we measure the extent to which surnames contain incremental information.

Two examples highlight the empirical advantages of the ICS measure. First, suppose that every individual had a unique surname. Then, by definition, surnames would contain zero information.

In this case $R_L^2 = 1$ but ICS=0 indicating, as it should, that surnames contain no information. Second, it is a fact that empirical surname distributions are highly skewed, with many surnames being shared by very few people and few surnames being shared by very many people. There are, therefore, many dummy variables in the D matrix from equation (3). The increase in R^2 attributable to D is likely to be large even if familial linkages contain zero information about economic outcomes. This is not the case for the ICS. The ICS is a more appropriate measure of informational content for skewed surname distributions.

In Section 7.2 we estimate the ICS using Spanish census data. We observe surnames directly and use educational attainment as a proxy for e_{is}. The vector X_{is} is comprised of data on age, gender, place of birth and, importantly, a measure of ethnicity. A preview of what we find will be useful as a frame-of-reference for our model, which we turn to next.

	R^2
Surname Dummies	0.3653
Fake Surname Dummies	0.3440
ICS	0.0213

Our baseline estimate of the Spanish ICS is 2.13%. At first blush this might seem small, especially when compared to the total predictable variation in the data. Our model will show (i) that 2.13% is *not* small once mapped into 'inheritance units' (the parameter ρ from equation (2)) and (ii) that the highly skewed nature of the surname distribution in western societies makes it natural that small ICS units have a large economic interpretation.

4 Model

The basic idea of our method is that the birth/death process for surnames results in a highly skewed surname distribution which, consequently, can reveal information about how economic outcomes are related across family members. The objective of this section is formalize this process in the simplest manner possible. Section 5 examines the model's robustness to various enrichments and extends it where necessary.

In Western societies the intergenerational transmission of surnames occurs primarily between fathers and sons. We therefore begin by ignoring females (this assumption is relaxed in Section 5.1). At date $t-1$ the population consists of N_{t-1} male individuals. Each individual reproduces

with probability q. Conditional on reproducing, an individual gives birth to m sons. Generations do not overlap. Fathers die after reproducing (or failing to reproduce). The expected growth rate of the population is $mq - 1$, which we assume to be zero.

Each of the N_{t-1} individuals is associated with one surname from the fixed, discrete set Ω. The typical element of Ω, denoted $s \in \Omega$, can take on one of S different values (so that $S \equiv \#\Omega$). The date $t - 1$ marginal distribution of surnames is $F_{t-1} : \Omega \to [0, 1]$. The number of *active* surnames at date $t - 1$, denoted $S_{t-1} < S$, is therefore equal to the number of strictly positive values of $F_{t-1}(s)$. If each individual has a unique surname then $S_{t-1} = N_{t-1}$. Otherwise $S_{t-1} < N_{t-1}$. In reality — and in our model — N_{t-1} is far greater than S_{t-1}. The initial distribution is denoted F_0.

The surname distribution evolves from F_{t-1} to F_t according to a birth/death process. The death of a surname occurs if all fathers possessing that name bear zero offspring.[3] Birth occurs via *mutation*: a son acquiring a different (typically new) surname than his father. Mutations are a necessary ingredient of our methodology. As we show in Section 4.1, without them the surname distribution would neither be informative nor would it resemble the highly skewed nature of observed surname distributions. Note that there is a certain irony here. Mutations seemingly frustrate the surname researcher: they 'destroy' intergenerational linkages. Yet without them the surname researcher would eventually be out of business.

Formally, mutation occurs as follows. At date t each existing name, $s : F_{t-1}(s) > 0$, will have vanished, so that $F_t(s) = 0$, if all fathers possessing that name at $(t - 1)$ bear zero offspring. This occurs with probability $(1 - q)^{N_s}$ where $N_s = F_{t-1}(s)N_{t-1}$, the number of fathers with name s. A surviving surname matches that of the father with probability $(1 - \mu)$ and *mutates* with probability μ. A mutated surname is simply a new name, $s \in \Omega$, chosen randomly.

Economic well-being is passed from fathers to sons according to equations (1)–(2) with $\gamma = 0$. That is, we ignore cross-sectional variation in the X_{is} directions so that individuals differ only in terms of surname and economic inheritance, y_{is}. Economic inheritance and well-being are therefore the same thing, $e_{is} = y_{is}$, which we refer to as *income* for simplicity. Rewriting equation (2), we have that the income of individual i with surname s at date t is determined by

$$y_{ist} = \rho \, y_{ip,t-1} + \varepsilon_{ist} \, , \tag{6}$$

---

[3] The death of surname s can also occur if all sons of all fathers with surname s mutate their names. Quantitatively, however, the likelihood of this happening is dwarfed by the likelihood that these fathers simply give birth to zero sons.

where $y_{ip,t-1}$ is individual i's father's income, one generation removed. By definition the surname associated with $y_{ip,t-1}$ is the same as that of y_{ist} (unless mutation occurs). Note that the mean of y_{ist} is zero, meaning that we are dealing with demeaned data relative to what is implicit in X_{is} from equation (1).

Siblings are individuals with y_{ist} and y_{jst} such that their surname s and parent p are the same (again, ignoring mutation). Identical surnames can also be associated with cousins, second cousins and so on. On the other hand, identical surnames can arise purely by chance, in the absence of any familial linkages. The smaller is $F_{t-1}(s)$, the less likely this is (*e.g.*, if $N_s = 1$ it is impossible).

The sense in which ρ relates to families and inheritance is manifest in the the distinction between the conditional and the unconditional variance of y_{ist}. For example, the cross-sectional variance between siblings is equal to the conditional variance, V_ε. For cousins it is $V_\varepsilon(1 + \rho^2)$. For the entire population it coincides with the unconditional variance, $V_\varepsilon/(1 - \rho^2)$. A larger inheritance parameter, ρ, therefore implies lower cross-sectional variance between family members *relative to* overall cross-sectional variance. The larger is ρ the larger will be the tendency for a surname to link two people with similar incomes, *relative to* two people randomly chosen from the population with, typically, different surnames.

This completes the description of the model. In subsequent sections of the paper we provide extensions that incorporate (i) variation across income groups in fertility behavior and surname mutation rates, (ii) females and assortative mating and (iii) ethnicity.

4.1 Analysis

We now discuss the key features of our model. Some features can be characterized analytically while for others we must rely on simulation-based evidence. For the simulations we use the following baseline parameter values. First, we abstract from growth so that the expected population growth rate, $mq - 1$ is zero. To achieve this, we set the reproduction probability to $q = 1/2$ and the number of offspring to $m = 2$.

Second, we choose the conditional variance, $V_\varepsilon = 1$, and the mutation rate, $\mu = 0.02$ Third, the initial number of individuals, N_0, is set to 1 million and the initial surname and income distributions are uniform. Finally, we vary the inheritance parameter, ρ, from 0.05 to 0.95. Whenever appropriate we examine the sensitivity of our results to departures from these baseline parameter

values.

The most important feature of our model is that skewness in the surname distribution gives rise to the informational content of surnames, and that this informational content is increasing in the inheritance parameter, ρ. We demonstrate this in the following sequence of properties.

Property 1 : Random walk behavior

Suppose that there is no surname mutation, $\mu = 0$, and the expected growth rate of the population is zero, $mq = 1$. Then the number of individuals with surname $s \in \Omega$ follows a driftless random walk with an absorbing barrier at zero.

The proof is in Appendix A. It is a simple consequence of the fact that the number of individuals with a given surname is a binomial random variable. Its importance is that it tells us that mutation is *necessary* for surnames to be informative. This is because a driftless random walk will, given enough time, visit all parts of its sample space. Eventually, therefore, all but one (non-informative) surname will disappear prior to the disappearance of the population (which, with $mq = 1$, must also eventually happen).

Next we demonstrate the way in which mutation generates skewness and thus informativeness. To do so we work with the ordered frequency distribution, denoted $G_t : [1, 2, \ldots, S] \to [0, 1]$. This distribution simply provides the relative frequency of the most common surname, the second most common surname, and so on. The long-run distribution associated with $G_t(k)$ is denoted $G(k)$, for $k = 1, 2, \ldots, S$.[4]

Property 2 : Skewed surname distribution

Given zero expected population growth, $mq = 1$, and a mutation rate, $0 < \mu < 1$, then for any initial distribution, $F_0(s)$ (and the associated $G_0(k)$), there exists a $k > 0$ such that, for all $t > k$, the distributions $G_t(k)$ display three key properties: (i) they are highly skewed, (ii) the number of individuals per surname is a constant, (iii) the Gini coefficient is a constant.

Figure 1 plots time-series, $t = 1$ to $t = 400$, of the number of individuals per surname and the Gini coefficient of the distributions $G_t(k)$. In each graph there are four time series, each one

---

[4]Formally, order the elements of Ω (arbitrarily) so that we can write $\Omega = \{s_1, s_2, \ldots, s_S\}$. Define the ranking function $\mathcal{O}_t : \Omega \to [1, 2, \ldots, S]$ as that which ranks each surname according to its commonality so that, for each $k = 1, 2, \ldots, S$, $F_t(\mathcal{O}_t^{-1}(k)) \geq F_t(\mathcal{O}_t^{-1}(k + j))$ for all $j > 0$ (ties are randomly allocated). $G_t(k)$, for $k = 1, 2, \ldots, S$, is then $G_t(k) \equiv F_t(\mathcal{O}_t^{-1}(k))$. The long-run distribution is then defined as $G : [1, 2, \ldots, S] \to [0, 1]$ such that, for $k = 1, 2, \ldots, S$, $G(k) = \lim_{t \to \infty} E[F_t(\mathcal{O}_t^{-1}(k))]$. Note that, since $F_t(s)$ is necessarily random for all t, we define $G(k)$ as the *expected* fraction of the population associated with the kth most popular surname.

corresponding to a different initial condition for the number of surnames (described in the caption). These moments of the distribution have clearly converged, thus validating Property 2. Since the Pareto distribution is completely characterized by these two moments, it seems likely that $G_t(k)$, also plotted in the figure for $t > k$ (along with the associated Lorenz curve), is a Pareto distribution. For our purposes, however, the exact form of $G_t(k)$ is not critical. What is critical is the behavior of the ICS, discussed below.

[Figure 1 about here.]

The skewness in Figure 1 is what drives our methodology. To understand why, consider first the names that occur with a high frequency. Since the income process in equation (6) is stationary, the cross-sectional distribution among these names is very similar to that of the overall population. These names therefore cannot be informative. In contrast, consider the very infrequent names. Many of them derive from recent mutations. They are newly created names, or the names of sons of fathers with newly created names, or grandsons, and so on. These names are *markers* that are likely to identify people with familial linkages.[5] If inheritance is important, so that these familial linkages connect people with relatively similar incomes, then the markers must make the same connections and, thus, be informative for income.

To make this clearer still, consider the evolution of the surname distribution versus the income distribution. They are (to this point) independent of one another. The frequency of a surname *cannot* be informative for income, in-and-of-itself. That is, what is *not* going on is that 'rich people have uncommon surnames.' What *is* going on is that low-frequency markers are indicative of familial linkages and high frequency markers are not. This is just as true for the rich as it is for the poor. In Section 5.3 we relax the independence assumption by allowing fertility rates to depend on income. For now, however, independence is useful in providing sharpness for the central mechanism of our method.

We now turn to our main result, the behavior of the ICS. The ICS is a moment of the *joint* distribution of surnames and income. Even though the two are independent of one another, the ICS connects them and reveals information about the latter based on the markers inherent in the

---

[5]For the same reason, surnames play an important role in the field of population genetics. The connection actually goes even farther. In our model, surnames are innocuous markers. They have no direct effect on income. Mitochondrial DNA, or the male Y-chromosome, is analogous. It does not code for any known protein and has no effect on the differential survival or reproductive chances of the individual receiving it. Nevertheless, it's a useful marker that allows researchers to make familial linkages.

former.

Property 3 : ICS and the importance of inheritance

Under the conditions of Property 2 the ICS from equation (5) is approximately constant for all $t > k$. Moreover, for any $t > k$, the ICS is monotonically increasing in the value of the inheritance parameter ρ.

The proof is in Appendix A. The monotonicity result is analytic whereas the constancy result is shown via simulation.

Figure 2 plots the ICS against ρ for our baseline parameter values. Aside from confirming the monotonicity property of Property 3, what's quite striking is the level and the convexity. Relatively small values for the ICS are associated with moderately large values of ρ and only for very high values of ρ do we see ICS values above, say, 10%. Again, echoing comments made above, this is necessarily the case given that only the rare surnames can be informative.

[Figure 2 about here.]

To summarize, the results of this section are as follows. First, without mutations the number of surnames will tend to become small, with each name conveying very little information about familial linkages and, therefore, inheritance. Second, mutations provide a countervailing force, allowing many rare surnames to have informational content. Finally, the informational content of these rare surnames — the ICS from equation (5) — is, *ceteris paribus*, monotonically increasing in the magnitude of the inheritance parameter, ρ. This is what allows us to identify the magnitude of ρ from data on the joint distribution of surnames and economic outcomes. In our Online Appendix we show numerically that these results are robust to different parameter values for the conditional variance of income, the mutation rate and family size.[6]

5   Extending the model

Our baseline model, above, captures the essence of what we are after. Nevertheless, it abstracts from some things that are important for the joint distribution of surnames and income. Most obvious is ethnicity. Less obvious, especially in terms of how it works, is gender and assortative

---

[6]The URL for our Online Appendix is `http://www.sevirodriguezmora.com/grt/`

mating. Finally, there is the possibility of dependence between fertility and income, something we've ruled out so far. In the empirical implementation of our model we address each of these things. Therefore, before getting to the data, we extend the model to provide guidance.

5.1   Gender and Assortative Mating

Assortative mating refers to the tendency of people with similar characteristics to marry each other.[7] At first blush, it may seem intuitive that assortative mating can give rise to the ICS because, ostensibly, it can generate "organization" in the distribution of surnames. If, for instance, today's rich and poor have distinct surnames, and if the rich marry the rich and the poor marry the poor, then one might think that the rich and poor surnames will remain distinct among future generations, thus generating informativeness. One might apply a similar argument to ethnically-motivated assortative mating. In either case, this intuition is deeply misleading. This is because the degree of assortative mating does not have any *direct* effect on the *marginal* distribution of surnames in the population.[8] The reason is simple.[9] Surnames are passed along the male lineage. For surname determination, it does not matter *why* one's father married one's mother, all that matters is one's father's name. It is as if females had no surnames.

What assortative mating *does* matter for is the *joint* distribution of surnames and characteristics. If assortative mating affects the income and/or ethnicity of a family's children and their descendants, *and* if surnames are informative for these things, then more assortative mating can give rise to more ICS. In the language of our model, if assortative mating increases inheritance between fathers and sons, ρ, then it will also increase the ICS — a moment, recall, of the *joint* distribution of surnames and income — because the ICS is monotonically increasing in ρ. The overall point is that the primary effect of assortative mating is not its effect on the surname distribution, but instead on things that the surname distribution can inform us about.

We now adapt the model of Section 4 to incorporate gender and mating. There is a *continuum* of males and females who form households and bear offspring. Fertility is (unlike above) determin-istic. Each household has one male child and one female child. Expanding on the notation above

---

[7] The existing literature on mobility that incorporates assortative mating includes Lam and Schoeni (1993), Chadwick and Solon (2002), Ermisch, Francesconi, and Siedler (2006) and Holmlund (2006).

[8] One exception is if assortative mating affects fertility. For example, if households formed by two wealthy spouses have different fertility rates for males than do households formed by two poor spouses, then the act of mating assortatively will change the marginal distribution (and the joint distribution) of surnames in the next generation.

[9] We are thankful to Melvin Coles for this insight.

(Equation (6)), y_{ist}^m and y_{ist}^f denote the incomes of male and female children who inhabit household i with paternal surname s at date t. This household was formed at date $t-1$. Its children's incomes arise as

$$y_{ist}^m = r z_{ip,t-1} + e_{ist}^m \quad ; \quad y_{ist}^f = r z_{ip,t-1} + e_{ist}^f \quad , \tag{7}$$

where the e innovations are *i.i.d.* $N(0, V_e)$, $r \in (0,1)$ is a *household* inheritance parameter, and z is average parental income: $z_{ip,t-1} \equiv (y_{ip,t-1}^m + y_{ip,t-1}^f)/2$. As above, subscript p denotes 'parent of the household' and notation for parental surnames is implicit; the father's surname is s and the mother's is irrelevant.

Given a cross-sectional distribution for $z_{ip,t-1}$, Equation (7) determines the cross-sectional distribution of (date t) male and female income that will then determine *parental* income, z_{ipt}, for the households of the next generation. This is where mating comes in. The incomes of the current generation's male children become that of the next generation's fathers: $y_{ipt}^m = y_{ist}^m$. Each father forms a household with a female who becomes a mother. The income of the mother is described by a *mating technology*, a function $f(y^m, u)$ that combines each father's income, y_{ipt}^m, with a *mating shock*, u_{ipt}, to assign to each father a spousal income, y_{ipt}^f, such that the distribution implied by $f(y^m, u)$ coincides with that implied by the inheritance process (7) for the population of female children at date t. The function we use is

$$y_{ipt}^f = \lambda y_{ipt}^m + u_{ipt} \quad ; \quad u_{ipt} \sim N(0, V_u) \quad , \tag{8}$$

where $\lambda \in (0,1)$ — the correlation between spousal incomes — is the degree of assortative mating. Note that Equation (8) is silent on the particular assignment mechanism that 'mates' the distributions of y_{ist}^m and y_{ist}^f from Equation (7). For our purposes, it is sufficient to simply form a set of ordered pairs, (y_{ist}^m, y_{ist}^f), that satisfy two properties: (i) they capture the notion of assortative mating that we are interested in, and (ii) they are consistent with the distributions implied by the inheritance processes (7). Examples of more fully-articulated assignment mechanisms are in Becker (1973), Gavilán (2011), Kremer and Maskin (1995), Marimon and Zilibotti (1999) and Shimer and Smith (2000).

In Appendix A we show that there exists a stationary distribution for average income, z. Inspection of the inheritance processes, Equation (7), then implies that, since $r < 1$, the stationary

17

distributions of male and female income must be the same,

$$y_{ist}^m \sim N\left(0, V_y\right) ; \quad y_{ist}^f \sim N\left(0, V_y\right) \quad , \tag{9}$$

where V_y is a unique function of the model's structural parameters, r, V_e and λ. It is given in Appendix A. Applying Conditions (9) to the mating rule, Equation (8), implies that

$$V_u = (1 - \lambda^2)V_y \ .$$

This is the condition that guarantees that the inheritance and mating rules imply the same cross-sectional distribution for female income.

Note that the inheritance parameter r from equation (7) relates male children's income to the income of their parents' *household*. The model of Section 4 and most of our empirical work, in contrast, refer to the correlation between children and their *father*. This is because surnames are passed along only the male lineage. Therefore, in order to understand how assortative mating affects the ICS we must describe how the parameters r, V_e and λ are manifest in both the variance of the income V_y and the parameter ρ from the following expression:

$$y_{ist}^m = \rho y_{ip,t-1}^m + w_{ist}^m \quad , \tag{10}$$

where the variance of w is denoted V_w. This equation links the income of sons to their fathers, a relationship that depends on both the mating process, (8), and the household-level inheritance process (7).

Note that, for issues of intergenerational mobility, the appropriate measure of inheritance is ρ and not r. This is because ρ associates comparable variables — the incomes of children with their father — whereas r from equation (7) does not. The latter associates the income of one individual with the consolidated income of their childhood household, something that arises from the noisy lottery of mating. One could, alternatively, use an analogous parameter that associates the consolidated income of each household with the consolidated income of that household's children's households. Indeed, a number of existing studies on mobility and assortative mating do just this. We choose to focus on ρ from equation (10) because (i) it is perfectly coherent, (ii) it is the parameter that estimated in the most of the existing literature on mobility, and (iii) it is tightly

linked to the process of surname diffusion.

In appendix A we prove the following property:

Property 4 *There exists a unique stationary distribution for y_{ist}^m and y_{ist}^f that is characterized by*

$$\begin{aligned} \rho &= \frac{r(1+\lambda)}{2} \\ V_w &= V_e\left(1 + \frac{r^2(1-\lambda)}{4\lambda}\right) \\ V_y &= \frac{V_e}{\lambda(1+\lambda)} \end{aligned}$$

A larger degree of assortative mating — as measured by a larger value for λ, the correlation of spousal income — thus translates into a larger value of ρ. Stronger assortative mating implies less intergenerational mobility in the population of fathers and sons. This is true even if the correlation between the income of sons and the joint income of their parents, r, is held constant. The intuition is straightforward. More assortative mating implies that the father's income is more informative for the income of the mother. Both father and mother contribute to the characteristics of their son. Thus, the more the income of the father explains the income of the mother, the more it must explain the income of his son. Stronger assortative mating translates into lower intergenerational mobility.

To summarize, surnames are passed exclusively along the male line. They do not provide any *direct* information about the mother. Any information that is *indirectly* associated with the mother must arise because the characteristics of the father are correlated with those of the mother. This is the mechanism through which assortative mating can affect the ICS. In the language of our model, the ICS depends *only* on the correlation and conditional variance of the incomes of fathers and sons: the parameters ρ and V_w, respectively. But assortative mating affects ρ and, therefore, it affects the ICS. This lends valuable interpretation to our empirical findings in Section 10.

5.2 Ethnicity

We think of the shocks ε_{ist} — defined above in Equation (6) — as being akin to measurable *productivity*. *Ethnicity* is something different. We define it broadly, as anything that is inheritable above and beyond our notion of productivity. Skin color, caste, geographic origin and some aspects of language are all examples of what we mean by ethnicity. In this section we consider the

implications of ethnicity having an effect on income. This could be for reasons of discrimination, regulation, asymmetric information, or many other things. We are silent on why ethnicity matters and focus only on describing its effects.

How is ethnicity related to surnames? In two ways. First, because surnames and ethnicity are passed along the male lineage, the former can be informative for the latter and, thus, for any associated economic effects. Second, if there is assortative mating in the ethnic dimension, then surnames can also be informative for the mother's ethnicity and, thus, can be informative for income for the reasons outlined above. We discuss each mechanism in turn.

In order to understand the first mechanism, consider again (as in Section 4) a world with only males. The difference is that now, in addition to surname and income, each male passes to his sons a permanent ethnic characteristic. Appending Equation (6), this characteristic affects income in the form of a "fixed effect," α_{eth}:

$$y_{ist} = \alpha_{eth} + \rho\, y_{ip,t-1} + \varepsilon_{ist} \ . \tag{11}$$

For simplicity assume that there are only two ethnic groups, red and blue (r and b) with $\alpha_r > \alpha_b$. Ethnicity and surname are passed along together. If there is an initial difference in the surname distribution of both groups, then future surnames are going to be informative on the ethnicity of their holders and, via this mechanism, on their holder's incomes. A surname that indicates that you are likely to be b also indicates that you are likely to be poor, as you are likely to have a low α.

Perhaps this is obvious. What is less obvious is that surnames must (eventually) reveal ethnicity *even if there are no ethnic differences in the initial distribution of surnames.* This is because the surname birth/death process is independent across ethnic groups. A surname mutation among the r group will generate a new name which, until the name dies-off, will only be associated with the r ethnicity. A surname death among the b group will leave relatively more r ethnic-group individuals with this name, thus increasing informativeness about ethnicity. The independent birth/death process will lead the r and b surname distributions to drift apart over time. Eventually an individual's surname will necessarily be informative on whether his ancestors were of the r or b ethnicity. If ethnicity is related to other characteristics like income, then surnames will also be informative on these characteristics, even though they may not have to begin with.

Consider next the second channel mentioned above: assortative mating. Females, of course, play an important role in determining the ethnicity of a household's children. Keeping in mind that surnames capture ethnicity *only insofar* as it is transmitted across the male lineage, it becomes clear that assortative mating is pivotal. It is the only way with which a mother's ethnicity can be correlated with her children's surname. Consider, for example, Judaism in which (ignoring conversion) ethnicity is *solely* passed along the maternal line. Absent assortative mating — *e.g.*, if Jewish women marry men randomly drawn from the entire population of males — surnames must eventually become uncorrelated with Jewishness of their holders. On the other hand, if Jewish women marry only Jewish men, then the surnames of Jews will become increasingly distinct from those of gentiles, owing both to the initial distribution of Jewish male surnames and to the surname birth/death process described previously.

This mechanism applies to virtually any other ethnicity-related characteristic. Since females are in almost all cases fundamental for the inheritance of ethnic characteristics, assortative mating and the degree of ethnic information contained in surnames go hand-in-hand. In the empirical implementation of our model, below, we take great care to control for such effects.

5.3   Surname Frequency

Our final model-extension involves relaxing a key assumption of Section 4, that the surname and income distributions are independent of one another. We now consider the possibility that fertility — that which drives dynamics in the surname distribution — may be related to income. If it is, then the frequency of an individual surname, $G_t(k)$, may be informative for income, in-and-of-itself. In this section we ask if this matters for our main results and if our model predicts any sort of systematic relationship between surname frequency, income and inheritance.

We build dependence between the surname and income distributions as follows. Birth rates, q, the number of sons, m, and the surname mutation rates, μ, are now allowed to be income-dependent: $\{q_r, q_m, q_p\}$, $\{m_r, m_m, m_p\}$ and $\{\mu_r, \mu_m, \mu_p\}$. Subscripts r and p ('rich' and 'poor') denote the upper and lower 20% of the income distribution, and m ('middle class') denotes the 60% in between. Population growth remains at zero, implying that $q_r m_r/5 + 3q_m m_m/5 + q_p m_p/5 = 1$. Respecting this constraint, the expected number of children, $q_j m_j$ can differ across income groups. In our Online Appendix (http://www.sevirodriguezmora.com/grt/) we use simulation

21

to demonstrate the following property.

Property 5 *If the fertility parameters and or the mutation rate depend on the position of the individual in the income distribution, then (i) surname frequency is informative for income, in-and-of-itself, and the sign and magnitude of the relationship depends on the specific parameter values for q, m and μ, (ii) the relationship between frequency and the inheritance parameter ρ is ambiguous, depending on q, m and μ, (iii) irrespective of parameter values the ICS is monotonically increasing in ρ.*

Elaboration and intuition are provided in our Online Appendix, where we also report the associated empirical evidence. We find that frequency and educational attainment are indeed related, albeit weakly.

The bottom line is that surname frequency is not useful for understanding mobility because the underlying cause of its correlation with economic outcomes is ambiguous and difficult to distinguish from ρ. Nevertheless, the utility of the ICS remains. Item (iii) tells us that, irrespective of the informational content of surname frequency, we can identify the degree of inheritance by looking at the ICS alone.

6 Data

We use data from two sources, the 2001 Spanish census and the 2004 Spanish telephone directory. From the census we have individual-level data from the Catalonian region of Spain on surnames, education and several other variables. From the telephone directory, obtained from Infobel, we have surname data. We describe the data and its uses below. First, however, it is important to understand how Spaniards name themselves and how this relates to our methodology.

6.1 Spanish Surnames

Spanish people have two surnames, 'first' and 'second.' The first is the first surname of their father and the second is the first surname of their mother. First surnames, therefore, are passed between generations in *exactly* the same manner as with the (traditional) Anglo-Saxon convention. Our methodology is based, primarily, on first surnames. It can be used in exactly the same way for males in Spanish societies as in many other Western societies.

This being said, the Spanish naming convention does offer additional information, information that we make use of. Unlike the Anglo convention, each male is connected with his mother. In addition, because females do not change their surnames upon getting married, each female is connected to her father. Note that this has no bearing on the evolution of the paternal lineage. The maternal surname vanishes after two generations.

We make use of this extra information in a number of ways. First, we use the second surname as a control for ethnicity, leaving the first as an indicator of the importance of familial linkages. Second, we use the combination of the two surnames to identify siblings and, thus, be more precise about familial linkages. That is, while two people holding as first surname "Rodríguez" are unlikely to be related, as are two people whose second surname is "García", two people named "Rodríguez García" are more likely to be related (as siblings). The likelihood grows as the frequency of the names decreases. Third, we use the combination of the two names to identify the strength of 'assortative mating' and its importance for economic inheritance. Fourth, we can use females in our analysis because their surnames are a link to their bloodline, and that of their siblings, not the bloodline of their husband's fathers.

One other distinguishing characteristic is the mutation rate of Spanish surnames. Compared to many other societies it's quite low. This is partly a consequence of Spanish orthographic rules and phonetics make that reproduction errors relatively unlikely.[10] It is also a consequence of societal norms and a complicated bureaucratic procedures that make it quite difficult to change one's name.

6.2 Census Data

The Census Data for Catalonia covers the entire population of 6,343,110 individuals. To our knowledge ours is the first paper to make use of the surnames from such an extensive census database. Our data consist of the two surnames of each individual, demographic characteristics (age, education, gender, marital status, place of birth, place of residence), as well as employment status, level of proficiency in Catalan language and several housing characteristics (tenancy, size, inheritance, availability of a second house). Other data is unavailable, for privacy reasons. For instance, we do not have information on explicit familial linkages, such as who a given person's

---

[10]For example, there is only one Spanish spelling of 'Rodríguez' — meaning "son of Rodrigo"; "Rodrígues" is a Portugese surname and spelling — while there are multiple spellings of Johnson, 'son of John'. This is well known and is one reason that surname-based studies on genetic inbreeding are often done in Spanish speaking countries (see Lasker (1985)).

father and mother are. We have some household data, such as the number of members of a given person's household, but we're unable to determine who these household members are. This does not, however, limit our methodology which requires only surnames and measures of economic well-being.

In Spain, census data do not include information on wealth or income. We therefore use years of education as our measure of economic well-being, y from equation (6). We eliminate individuals living in 'collective households' because the census has no educational information on them. We also eliminate individuals from whom the first or second surname is missing. This leaves us with 6,123,909 individuals. For our analysis, we consider only individuals aged 25 and above, who are likely to have finished full time education. We include only Spanish-born, Spanish citizens so as to mitigate the extent to which surnames are informative because they distinguish an immigrant who is likely to have relatively low education. Finally, we exclude individuals with a unique first surname because such names cannot, by definition, provide familial linkages with other individuals. We refer to our dataset, after all of these eliminations, as the 'complete population.'

6.3   Telephone Directory

Immigration — actually the lack thereof — plays an important role. Prior to 2001, almost all migration involving Spain was either intra-national or was emigration, from Spain to the rest of the world. This makes controlling for ethnicity relatively easy because it is the same as controlling for regional ancestry. This is what makes the telephone book useful. We use it to both compute the distribution of surnames in the whole of Spain and obtain a measure of how 'Catalan' a given name is. The latter provides a valuable control for the possibility that Catalan origin, and not familial linkages, is underlying the informational content of surnames.

We obtain the Spanish telephone directory from a commercial source. There are roughly 14 million households in Spain and the directory contains surname information on roughly 11.4 million private, fixed telephone lines. Mobile phones, which are not included, obviously account for the majority of the difference. We have no reason to believe that the surname distribution differs across fixed versus mobile lines and are therefore confident that the absence of the surnames of mobile-only households does not affect our results.

6.4 The Surname Distribution

In this section we describe the surname distribution associated with the census and with the telephone directory. For comparability reasons, we include only those individuals who are listed with both a first and a second surname.

[Figure 3 about here.]

Figure 3(a) plots the commonality-ordered frequency distribution of the first surname (the empirical counterpart to G_t from Property 2 in Section 4.1). The distribution is very skewed. There exist a large number of low-frequency surnames and the few most frequent surnames represent a large percentage of the population. The 10 most popular names cover roughly 11 percent of the population.

[Table 1 about here.]

Figure 3(b) plots the Lorenz curve for the Spanish and Catalan surname distributions, using both the telephone book and census information for the latter. Table 1 reports the relevant statistics for the three distributions. Notice that the number of people per surname is larger in the whole of Spain than in Catalonia. This is probably because Catalonia is a net receiver of immigration and because the Catalan language has, historically, had less orthographic rigidity than the Spanish language. There are also more surnames in the census than telephone directory, no doubt in part because each household typically has just one telephone line.

The data can give us some idea of the magnitude of the flows in and out of the surname pool. Regarding the outflow, the fraction of women aged 50 or above who have had zero male children is 0.33. From this we calculate that the expected number of surnames that vanish is in the range 100 to 110 per year.[11]

---

[11]If males and females have equal probabilities of having no male children, then, given the distribution of surnames, the expected number of names that would vanish each year would be 96. This is probably an underestimate for two reasons. First, it does not take into account women younger than 50 who die without having male children. Second, it is likely that the number of males having children is smaller than the number of females having children (*e.g.*, some males mating with more than one children–bearing women and others with none). We have no way of looking at this, but it is reasonable to assume that it doesn't amount to much. On one hand the life expectancy of women is very high in Spain, substantially more than 71 years at birth. Additionally the uneven distribution of the number child-bearing mating partners for males is likely to be less pronounced when one refers to surnames than when one refers to actual births. This is because children are likely to carry the surname of the husband of their mother, not necessarily of their father. We thank Namkee Ann for providing the data based on the 1991 Spanish Socio-Demographic Survey.

The inflow is more difficult to measure. It is comprised of immigrants into Catalonia with new names and mutations of existing names. We ignore the effect of immigration because we remove first-generation immigrants from our population and because it's difficult to know how likely it is that an immigrant's name is new. For example, a South-American immigrant is likely to have an Hispanic surname which, in some sense, represents a name that emigrated from Spain and is now returning. We therefore focus our measure of the inflow upon surname changes of existing individuals. In 2001, in all of Spain, there were 1,570 applications to change one's surname.[12] Of this, 1,426 were granted. Assuming that 2001 is a representative year, and a population of around forty million with a live expectancy of around 70 years, this amounts to a mutation rate of around 0.0025. Using the sample of all Spanish males living in Catalonia, this translates to 107 new surnames each year. Thus, our calculations suggest that, caveat the difficult-to-measure immigration effects, the flows in and out of the surname pool are of similar magnitudes.

7 Cross-Sectional Results

In this section we present results that are 'static' in nature. This means that we pool together individuals from different birth cohorts. We obtain cross-cohort *average* measures of the ICS. and we calibrate the model to obtain a comparable measure of intergenerational mobility. In Section 9, in contrast, we condition on birth cohort and obtain measures of how these things have changed over time. We begin by discussing how we deal with the issue of ethnicity.

7.1 How Catalan is a Surname?

As we saw in section 5.2, surnames carry information on both ethnicity and familial linkages. We seek to isolate the latter and therefore need to control for the former.

We control for non-Spanish ethnicity by simply excluding all individuals who are not Spanish-born, Spanish citizens. Because Spain (including Catalonia) saw very little foreign immigration prior to the nineties, this leaves us with a population that is primarily of Spanish origin. Among these people an important aspect of 'ethnicity' is *regional origin*: the region of Spain from which one's family originates.

---

[12]Data from the Office of Public Records (*Registro Civil*). We thank Manuel F. Bagüés for providing us with these data.

Regional origin in Spain is tightly connected to language. In Catalonia, native speakers of the Catalan language — 'Catalans' — have a strong tendency to be of Catalan regional origin. It is well known (and further verified below), that Catalans tend to have (i) surnames that are distinctively 'Catalan,' and (ii) relatively high education and economic well-being relative to those whose maternal language is Castillian Spanish. Surnames in Catalonia, therefore, are likely to be informative for reasons that go beyond familial linkages.[13] What we'd like to do is to eliminate this "ethnicity" component, thereby focusing attention on familial linkages.

Consider the following index of the *"Catalonianess"* of a particular surname, s.

$$CatalanDegree(s) = \frac{\text{Number of telephones under surname } s \text{ in Catalonia}}{\text{Number of telephones under surname } s \text{ in Spain}}$$

Taken literally, this ratio is an estimate of the probability that an individual with surname s has a telephone registered under their name in Catalonia. A little less literally, it is an estimate of the fraction of people with surname s that reside in Catalonia. For example, since the 2001 Catalan population is about 16% of the total Spanish population, then, if surnames were uniformly distributed throughout Spain, this ratio would be roughly 0.16. The extent to which it's higher (lower) indicates a concentration of people with surname s residing inside (outside) of Catalonia.

We will go one step further and interpret the $CatalanDegree(s)$ variable as a proxy for extent to which a person with surname s has Catalan regional origin. How good of a proxy is it? We know several things that can help us understand. First, a large percentage of intra-Spanish immigration into Catalonia occurred after 1955. Second, this immigration flow was large; without it Cabré (2004) estimates that the year 2000 population would have been 2.7 million instead of the actual value of roughly 6 million.

These facts tell us that, in the 2001 census, elderly people, and especially those born in Catalonia, are more likely to be of Catalan origin than younger people. If our $CatalanDegree$ variable is a good proxy for regional origin, it should therefore reflect this. Table 2 shows that it does. The overall average of $CatalanDegree$ in 0.35, whereas among people born prior to 1950, in Catalonia, the average is 0.57. Figure 4 elaborates. It plots the mean and standard deviation of

---

[13]The reasons range from obvious to controversial. An obvious one is the initial condition. It is well known that immigrants into Catalonia have been considerably less wealthy and less educated than the native population. Controversial reasons include the linguistic advantage that native Catalan speakers have in the educational system (see Aspachs-Bracons, Clots-Figueras, Costa-Font, and Masella (2008)), and a variety of forms of discrimination that non-catalan speakers may be subject to.

CatalanDegree for an overlapping sequence of 25-year age-cohorts.[14] The surname distribution in Catalonia has clearly become 'less Catalan' over time, as the immigration flows tell us it should if our proxy is a good one.

[Table 2 about here.]

[Figure 4 about here.]

As further support for the quality of our *CatalanDegree* proxy we run two probit regressions. In the first, the left-hand-side (LHS) variable takes value 1 if an individual has full knowledge of the Catalan language.[15] The right-hand-side (RHS) variables are, in column (1), individual-specific controls (gender, place of birth, age, regional dummies). In column (2) our *CatalanDegree* variable is added. We estimate a large, significant, positive probability. Figure 5(a) shows the estimated probability for the relevant range of the *CatalanDegree* variable.

The second regression asks how well *CatlanDegree* predicts immigration history. The LHS variable takes value 1 if an individual 50 years of age or older immigrated into Catalonia from elsewhere in Spain. Results are reported in Table 3(b) and Figure 5(b). The estimates are negative, large and significant and the pseudo-R^2 increases dramatically with the inclusion of *CatalanDegree*. People with lower *CatalanDegree* surnames are much more likely to be immigrants than those with higher *CatalanDegree* surnames.

[Table 3 about here.]

[Figure 5 about here.]

To sum up, *CatalanDegree* appears to be a good measure of Catalan regional origin. In what follows we'll use it in two different ways. First, in our ICS regressions we'll include on the RHS the *CatalanDegree* variable associated with an individual's *second* surname (their maternal surname). A dummy variable for their *first* surname will be included for measurement of the ICS (Section 3, equation (5)). Using the first and second surnames in this manner is meant to mitigate multicollinearity. Second, in Section 7.2, we consider sub-populations of regionally-homogeneous

---

[14]More specifically, the sequence starts with those aged 75-100 years old in 2001, then continues with those aged 70-95 years old, and so on, ending with the 25-50 year-old cohort.. The same sequence will be used in our dynamic, cohort-based analysis.

[15]The census question asks a resident if she speaks, reads and writes Catalan. Roughly 45% of the over-25 population responded in the affirmative.

groups. We calculate the geometric mean of the *CatalanDegree* for both the first and second surname and then order the population according to the associated values. We then identify upper 40% quantile as a homogeneous group of individuals with catalan regional origin.

7.2   The ICS in 2001

Table 4 reports a benchmark set of estimates of equations (3–4) and the associated ICS from equation (5). Column 1 begins by including only individual controls — dummy variables for gender, age and place of birth.[16] The adjusted R^2 is 0.3363. Column 2 adds our *CatalanDegree* variable. The coefficient is positive and highly significant.[17] It's also economically significant. The standard deviation of *CatalanDegree* is about 0.3. Therefore, the estimate of 1.692 translates into an additional 0.5 years of education for a one-standard-deviation increase in a surname's 'Catalonianess.' The mean and standard deviation of education are 8.0 and 4.7, respectively. So Catalan regional origin is associated with higher educational attainment equal to about 10% of the overall dispersion.

[Table 4 about here.]

Column 3 of Table 4 adds paternal surname dummies to the regression (recall that maternal surnames are used to define *CatalanDegree*). There are $38{,}024$ surnames, or roughly 113 people per surname. We note that (i) the surname dummies are jointly significant (given the large number of RHS variables involved this is not obvious in spite of the large population size), (ii) the coefficient of *CatalanDegree* is smaller but remains economically meaningful, with a one-standard-deviation increase translating to 4 extra months of education, and (iii) the R^2 increases to 0.3653. Surnames are thus informative. Knowledge of the *particular surname* of an individual is informative for predicting their educational attainment.

Column 4 replaces the actual surname dummies with 'fake' dummies as in equation (4) of Section 3. The fake dummies are not jointly significant and their presence increases the R^2 very little. The estimate of the *CatalanDegree* coefficient is largely unaffected. Our estimate of the

---

[16]We define place of birth differently for those born in Catalonia versus those born elsewhere in Spain. If Catalan-born, we use county dummies, otherwise we use Spanish province dummies. Catalan counties are administrative units somewhat smaller than a typical U.S. county. Spanish provinces are somewhat larger than a typical French *departament*. We also used, instead, town-of-birth dummies and found very similar results.

[17]The very small standard errors of our estimates are obviously due to a very large sample size. As a result, most of our discussion focuses upon the economic magnitude of the coefficients, not their statistical significance. One important exception is the joint significance of the actual versus fake surname dummies.

ICS from equation (5) is, therefore, 2.13%. Columns 5 and 6 are analogous to columns 3 and 4 except that the *CatalanDegree* variable is omitted. Since surnames now capture both Spanish regional origin and familial linkages, we expect the ICS to increase. It does, to 2.66%.

[Table 5 about here.]

Tables 5(a) and 5(b) repeat the exercise of table 4, but restricting the population to those born in Catalonia (table 5(a), immigrants are not included, even if their children are) and to those born in Catalonia before 1950 (table 5(b)). The results are very similar to our benchmark in Table 4, the only exceptions being that the R^2s are smaller and the ICS is slightly higher at 3.19% and 3.26%, respectively. Both are to be expected given that the population is more ethnically homogeneous.

One concern is that, in spite of our use of *CatalanDegree*, our results are dominated by ethnic and not familial linkages. To examine this, Table 5(c) restricts the sample to the 50% of the population who have surnames with the highest *CatalanDegree* (see Section 7.1 for the precise definition). The qualitative nature of our results is unaffected, with the ICS being just slightly higher at 2.43%. Note that, encouragingly, the ICS is basically unaffected by the inclusion of the *CatalanDegree* variable.

To summarize, our results show that surnames are informative for educational attainment. Part of this is because surnames are informative for Spanish regional origin. But an important part remains, even after controlling for regional origin in a variety of different ways. Our interpretation — that elucidated by our model — is that the surname partition is informative about familial linkages. We now present two additional pieces of evidence that support this interpretation.

7.3   Rare Surnames are More Informative

Our model predicts that, if inheritance is important, surnames will be informative because the partition of *rare* surnames should group together people with familial linkages. This then implies that the ICS should *increase* as we exclude common names. Checking this in the data provides a valuable check on our interpretation of our results.

[Table 6 about here.]

Table 6 repeats the exercise of subsection 7.2 but includes only the 50% of the population with the least frequent surnames. As our model predicts, the ICS increases, from 2.13% (Table 4) to

3.22%. Figure 6 provides additional evidence. It is based on a series of regressions, analogous to those in Table 4, that sequentially include people with more and more common surnames, as we move from left to right on the horizontal axis. The left vertical axis reports the ICS — the downward-sloping line — and the right axis reports the average number of individuals per surname in each sub-population.

Figure 6 provides strong evidence in favor of our model's interpretation of the ICS. Moving from least frequent names to the most frequent, the ICS monotonically falls by a factor of seven. This suggests that our controls for ethnicity are working and that our findings are driven by the informativeness of surnames for familial linkages.

[Figure 6 about here.]

Note a key distinction. Here, we have shown that as average surname frequency decreases, the informativeness of the *individual surnames* increases. This is to be distinguished from the informativeness of the *individual surname frequencies*. The latter would ask, for example, 'is someone with a rare surname likely to be highly educated?' This is *not* what we ask here. We do ask it in our our Online Appendix, where we find in the affirmative and provide discussion. For now, recall that Property 5 demonstrates that the important properties of the ICS are not affected by the informativeness (or lack thereof) of surname frequency.

7.4 Invented Catalonias

Our results should not be sensitive to any random (but sufficiently large) partitioning of the surnames set. One such partition is simply based on the alphabet. If, for example, we randomly assign letters of the alphabet to two groups, "first half" and "second half," the ICS should be unaffected. The same applies if we just pick the median letter in the alphabet and divide the population there. Table 7 does exactly this.

[Table 7 about here.]

The first column reports, for comparison purposes, the overall ICS from Table 4. The second and third report the same statistics but for two "invented Catalonias:" those from the first half of the alphabet and those from the second half.[18] As our model predicts, neither the R^2 of the

---

[18]We have done the experiment with other random groupings, and obtained the same result.

regressions nor the ICS change across the populations. This suggests that our findings are structural and that they depend on deeply rooted social and economic mechanisms.

8 Calibration

Here, we calibrate the model of Section 4 using the results of Section 7. The main task is to obtain a value of the inheritance parameter, ρ, that results in our model's ICS matching the value of 2.13% from Table 4 of Section 7.2.

We choose our model's parameter values as follows. As in Section 4 we abstract from population growth, setting the reproduction probability $q = 1/2$ and the number of sons $m = 2$. This also assumes no dependence between fertility and income/education. We set the mutation rate, μ equal to its sample counterpart from our data, 0.0025 (see Section 6.4). This generates a steady-state Gini coefficient of 0.82, which is maximal for our model, but less than the value of 0.903 that we observe in the data. Our fertility/mutation process is not sufficiently rich to account for the skewness observed in reality.

As in Section 6, we use educational attainment as a proxy for economic-well being. We assume that it has a continuous distribution and that it follows a Gaussian AR(1), as in Equation (6), but in logs instead of levels. We allow for a positive unconditional mean, θ and write $\log y_{ist} = (1-\rho)\theta + \rho \log y_{ip,t-1} + \varepsilon_{j,t+1}$, where $\varepsilon_{j,t+1} \sim \mathrm{N}(0, V_\varepsilon)$. The mean, θ is set to match our data at 9.18 years of education. The conditional variance, V_ε, is set so that, conditional on ρ, the unconditional variance of y_{ist} matches the sample value of 4.56.

Given these parameter values, we choose the number of individuals in the population to be 20,000 and the initial distribution of surnames and log-education to be uniform and Gaussian, respectively. The latter is set equal to the unconditional distribution of $\log y_{ist}$. We then fix a value of ρ and simulate the economy for 100 generations so that, in the sense of Property 2, convergence will have occurred. The 101st generation gives us one theoretical census. We use it to run the dummy-variable regressions from Section 3, Equations (3) and (4), and then we compute the model's ICS as in Equation (5). Thus, we have one ICS for each value of ρ, a mapping that we designate as $ICS(\rho)$.

[Figure 7 about here.]

32

Figure 7 graphs the function $ICS(\rho)$ for all allowable (positive) values of ρ. The calibrated value for ρ — that which generates an ICS of 2.13% — is 0.47. This result is of interest along several dimensions. First, it substantiates the point made at the end of Section 3; small 'incremental R^2' units for the ICS map into large 'inheritance units' for ρ. This is the nature of our method and it's a natural consequence of extracting information from a highly skewed surname distribution. Second, $\rho = 0.47$ is similar in magnitude to the analogous estimates — based on very different data and methodology — that we discuss in the literature review of Section 2. Finally, our calibrated value for ρ is a useful reference point for our next exercise, asking how the ICS has changed over time. Whereas 2.13% is based on the entire census, we estimate values of 1.98% and 3.50%, respectively, when we divide the census into old and young cohorts. These numbers map into (steady-state) inheritance units of $\rho = 0.41$ and $\rho = 0.55$. While it is unlikely that the distribution has transited from one steady state to another in just a few generations, we find these calibrated values informative nevertheless.

9 Dynamic, Cohort-Based Results

Our analysis to this point has treated the entire 2001 Catalan census as a single cross-section. The cross-section, of course, consists of individuals of different ages. We have dealt with education-related age effects using dummy variables. But we have not allowed the ICS to vary with age. The above estimates are age-averaged measures of the ICS. They are 'static' in the sense that they are incapable of saying anything about how the ICS and intergenerational mobility may have changed over time.

We turn now to a dynamic analysis. We partition the cross-section into birth cohorts and ask if the ICS varies across subpopulations of different ages. If it does, our model suggests that intergenerational mobility may have changed over time. The reasoning is as follows. Suppose that mobility has decreased, so that the value of ρ connecting the current young generation to their parents has increased. Our model's prediction is that, if the surname distribution of the parents is highly skewed (which it is), then the surnames of the young should be more informative than those of the old. The change in mobility should generate a change in the ICS. Why? For the same reasons as above. A higher value of ρ implies that that familial linkages will be more informative for economic status. A skewed surname distribution of parents means that the surname distribution

33

of children will be informative for familial linkages.

There are, of course, alternative interpretations for why the ICS may have changed over time. One possibility is a change in the birth/death process for surnames. We discount this possibility for one simple reason. Such a change must necessarily affect the ICS via its effect on the surname distribution. This effect will occur very *slowly*. Its immediate effect on the ICS, therefore, must be very small. A change in ρ, in contrast, will have an *immediate* impact on the ICS because the channel through which it works is not the slow-moving surname distribution. Our empirical robustness checks (based on siblings in Section 11 and based on assortative mating in Section 10) serve to bolster this point and provide reassurance that we are capturing changes in ρ.

Finally, before getting to the results, there is one last interpretive issue that needs to be understood. Since our data centers around educational attainment, one must be aware of the large, secular changes that occurred in Catalonia and Spain during the mid 20th century. Prior to the 1950's access to education was very limited. This was a consequence of both the general level of wealth and income as well as the lack of investment in public education. Starting in the late 1950's things began to change. Both the economy and the level of public education began to grow, particularly from 1975 onward. This shows up clearly in our data. Figure 8 demonstrates that the average years of education of the oldest individuals in 2001 is less than half that of the youngest. Variability around the average, in contrast, is more stable.

[Figure 8 about here.]

How is 'more education for all' associated with intergenerational mobility in educational attainment? Does an increase in publicly-provided education decrease the importance of how educated one's parents are? While the intuitive answer might be yes, we now show that it doesn't fit the facts. We go on to argue that such intuition confuses aggregate growth with cross-sectional mobility. Mobility is a *relative* concept. Aggregate growth is not. Further discussion is provided below, in Section 9.3.

9.1 Cohort-Level ICS

[Table 8 about here.]

Table 8(a) reports results for the same regressions as Table 4, except that the population is re-

34

stricted to those born *before* 1950.[19] The results are similar to those for the entire population. The adjusted R^2 of the regression with real surname dummies is 0.2533, compared to 0.2335 with fake dummies. The ICS is therefore 1.98%. Table 8(b) includes only those born *after* 1950. There are three notable results. First, the explanatory power of the regressors is much lower, generating an adjusted R^2 of 0.1534. Not surprisingly, gender and geographical location explain less of the variation in education in the post-war period, surely reflecting the more widespread access to education. Second, the parameter of *CatalanDegree* is substantially larger. Regional origin has become more important for determining educational outcomes. Figure 9(b) emphasizes this, plotting our estimates of *CatalanDegree* for the same rolling window of cohorts used in Figure 8.

Finally — and most importantly — the ICS is substantially higher for younger cohort. It is 3.5% for those born after 1950, a 75% increase over that associated with those born before 1950. Figure 9(a), again, shows results for finer partition of age cohorts. Our interpretation — which we go on to substantiate below — is that intergenerational mobility has decreased in Catalonia, even after controlling for a strong effect of the increased importance of ethnicity.

[Figure 9 about here.]

9.2 Robustness Checks

Decreasing intergenerational mobility in the face of the large increase in overall education is an important and provocative result. We now present a battery of robustness checks, basically replicating the checks that we undertook for the single cross-section in Section 7.

[Table 9 about here.]

Table 9 replicates the results of Table 8 but restricting the sample to the 50% of the population with the most Catalan surnames. As before, the ICS increases (from 2.60% to 4.27%), even though the other RHS variables have much less explanatory power for the young than for the old. Notice that this is a much more homogeneous group in the ethnic dimension.

[Table 10 about here.]

---

[19]The data include both people born in and outside of Catalonia. If we were to exclude the latter (as in Table 5(a)), we would include the children of the immigrants in the sample of 'young' but not their parents in the sample of the 'old'. Thus, to look at time trends could be misleading.

Table 6 and Figure 6 from Section 7.3 provided a powerful confirmation of our model by showing that the ICS is much larger when we consider only (relatively) rare surnames. Table 10 replicates the analysis for age-cohorts, excluding individuals with surnames that are in the upper 50% of the commonality distribution. Again, the ICS increases (from 3.52% to 5.05%) when we exclude names that, almost by definition, cannot be informative about familial linkages.

[Figure 10 about here.]

Figure 10 shows the same growth of the ICS using the same moving windows of 25 years as Figure 9 once we restrict the population to the individuals with high *CatalanDegree* (10(a)) and individuals with low surname frequencies (10(b)).

9.3 Discussion

Our cohort analysis arrives at three important findings. First, alongside the large increase in average educational attainment, there has been a decrease in the explanatory power of gender, age and place of birth. Second, ethnicity has become more important for educational attainment. Third, the component of the ICS that is unrelated to ethnicity has increased. We now discuss each finding in more detail.

Increase in educational attainment. To be female or to be born in a rural environment became less important for explaining the education of an individual. This can be observed in the decrease in the R^2 of the all the regressions when we use only the younger population. It is clear from column (1) in tables 8(b), 9(b) and 10(b) that the percentage of the total variance explained by the individual controls is much lower. Likely explanations are (i) the widespread increase in education attainment and investment in public education over the period, and (ii) the increase in the social and economic status of females. Thus, people from different locations and of different genders have more similar fates than those of previous generations.

Note that this does not refer to the parents of the individuals. It does not refer to the specifics of their upbringing or their parent's status. In our analysis this is reflected in the variance explained by surnames, both in *CatalanDegree* (as a proxy of ethnicity) and directly (as an approximation to family networks).

Increase in the importance of ethnic background. Our point estimates of the coefficients on the *CatalanDegree* variable have increased, indicating an increase in the importance of ethnicity. Note that this cannot be a direct consequence of the migration process. In our regressions we include not only the children of immigrants, but also the immigrants themselves. Thus, our result indicates that *CatalanDegree* is more important for measuring the education of the second than of the first generation of immigrants. It is not the case that it increases because there are more immigrants. Note also that this does not mean that low *CatalanDegree* individuals have obtained less education, but that their difference vis-a-vis high *CatalanDegree* individuals has increased. The increase in educational attainment has been large across-the-board for both ethnic groups, but has affected Catalan speakers more than non-Catalan speakers.

In Catalonia there are two main linguistic communities, Catalan and Castillian (Spanish), each representing roughly half the population. Catalan speakers have enjoyed substantially larger incomes and larger levels of educational attainment during the entire period of our study (this is true for both those born before and after 1950). Nevertheless before the late 1970's there did not exist any formal linguistic advantage toward Catalan speakers. The language of government, commerce and education was overwhelmingly Spanish. However, beginning in the late 1970's the increasing political power of Catalan nationalism has translated into a series of drastic legal and administrative reforms that have turned upside down the relative importance of both languages in society while changing only marginally its overall language composition.[20] For example, since the beginning of the 1980's all education is provided *exclusively* in Catalan in all public and practically all private schools. Catalan is now the sole language of the regional and municipal governments, and proficiency in Catalan has been the key requirement for working in public administration since the beginning of the 1980's. Further legal change has made Catalan an important (albeit perhaps not the main) business language.

Governmental and institutional changes in the use of Catalan are, at best, a partial explanation for what we find. A deep understanding of the increase in the value of ethnicity is beyond the scope of this paper. Note, however, that in Section 10 we do dig a little deeper and show that

---

[20]See Miley (2004) for a study of the politics of nationalism and language in Catalonia. The increasing power of Catalan Nationalism might be explained (i) by the larger levels of income and education of the Catalan speaking community and (ii) because Spanish electoral law has allowed Catalan nationalism to operate as a third party in Spanish politics, allowing it to obtain high leverage from its successive alliances with either left or right leaning governments. See also Aspachs-Bracons, Clots-Figueras, Costa-Font, and Masella (2008) for a study of the effects of linguistic legislation on the educational system on identity.

assortative mating seems to have increased in Catalonia, in part along ethnic lines. This makes ethnicity more inheritable and serves to magnify the effects of any institutional changes on the educational outcomes of the offspring of those who assortatively mate.

Decrease in intergenerational mobility. The increase in the ICS, even after controlling for ethnicity, is probably the most important of our empirical results. One possible explanation (*c.f.*, Checchi, Ichino, and Rustichini (1999), Checchi, Ichino, and Rustichini (1999)) is based on the increase in the provision of public education. The main idea is that public education is a subsidy, reducing the price of education for all. This induces an increase in the *average* level educational attainment. However, in the cross-section, not all individuals have the same demand for educational services. The children of the rich and/or more educated may demand more because of aptitude, motivation, incentive, role models, and so on. If so, they benefit disproportionately from the subsidy. Thus, even if the average individual gets more education, the fact that one's parents are rich or poor may be instrumental in determining where one falls in the distribution around the average.

Our methodology and data do not allow us to evaluate this explanation. We can, however, examine a second possibility. The decrease in mobility may be a result of an increase in the degree of assortative mating one generation before. On this we have direct empirical evidence that is quite conclusive. The Spanish naming convention, once again, proves helpful here. We turn to this next, using the framework developed above in Section 5.1.

10   Assortative Mating

We have seen that surnames contain information on two such characteristics: ethnicity and educational attainment. This, combined with the Spanish naming convention, allows us to obtain measurements of the level and change in ethnic/educational assortative mating in Catalonia. As we have seen in sections 5.1 and 5.2, an increase in the degree of assortative mating translates into an increase in the prevalence of inheritance, and of the ICS.

Our identification strategy is best illustrated with an example. The example emphasizes the ethnic dimension of assortative mating, but the logic applies equally to the educational dimension. The surname Casals is associated with a high value of our *CatalanDegree* variable. The same

applies to the surname Pujol. A person whose complete surname is "Casals Pujol" — a person whose father is Casals and mother is Pujol — is therefore almost certainly a person with two parents of Catalan regional origin. Ethnic assortative mating, then, can be measured by the incidence of such complete surnames relative to those that are more ethnically-heterogeneous. The measurement is simple correlation between the ethnicity index of each person's first and second surname.

Note that this measurement applies to each individual's *parents*, not to each individual's spouse. That is, if we find evidence of increased assortative mating among the 25-30 year-old cohort in the 2001 Catalan census, this means (very roughly) that the 50-55 old cohort exhibited more assortative mating than those one generation older.

[Table 11 about here.]

Table 11 contains our results. The data are constructed by first associating to each surname two characteristics: the average level of education and the average value of $CatalanDegree$, where the average is taken across all individuals with that particular surname. We then run two sets of regressions, one for each characteristic. The LHS variable is each individual's first surname's characteristic and the RHS variables are the set of controls used above along with their second surname's characteristic. Table 11(a) reports results for the education characteristic, again partitioning the population into those born before and after 1950. We see that the correlation between the educational dimension of first and second surnames increases from 0.170 for the old cohort to 0.303 for the young. Educational assortative mating seems to have increased. Table 11(b) reports the analogous measurement for ethnicity. The correlation also increases, from 0.217 to 0.328. The parents of younger cohorts seem more likely to have married within their ethnic background than the parents of the older cohorts.

[Figure 11 about here.]

A graphical representation is given in Figure 11. We plot the value of the parameters for the regression of education (Figure 11(a)) and for the regression on $CatalanDegree$ (Figure 11(b)) for the moving window of cohorts described above. They are both clearly increasing, and education has a timing that resembles the timing of the increase in the ICS.

[Figure 12 about here.]

In order to make sure that our results are not driven by ethnicity we run the same regressions on ethnically homogeneous populations (Figure 12(a)) and with very infrequent surnames (Figure 12(b)). The results are qualitatively identical.

To summarize, we have found evidence that intergenerational mobility in educational attainment has decreased in Catalonia in the 20th century. One possible explanation is that assortative mating has increased. Surname data is consistent with explanation, suggesting an increase in the likelihood that people mate with others of similar educational levels and ethnic backgrounds.[21]

11   Methodological Robustness: Sibling Correlations

Our method is based on the *surname* partition of the population. If, instead, we partition the population into groups of *siblings* then — if our method is valid — we should find that the ICS is higher relative to the surname partition. This is because surnames are a noisy indicator of familial linkages whereas the sibling relationship is not. In this section we ask if the ICS is indeed higher for the sibling partition. If it is, this serves as a powerful robustness check on our method. This is because, while we *use* surnames to form the sibling partition, the basic data that we then average across to form the ICS is very different than its surname-based counterpart. Moreover, a sibling-based analysis provides an important link to related work (cited in Section 2) that uses data on siblings exclusively.

The sibling idea is simple. First, recall (from Section 6.1) that all Spaniards have two surnames, the first from their father and the second from their mother. Thus, all siblings (irrespective of gender and marital status) share not one but two surnames, as well as their ordering. This allows us to construct a partition of the population which *we know* will tend to group together individuals with immediate familial linkages.

Define the "complete surname" for an individual to be their two surnames, in order. That is, if a person's father and mother are named Fernández and Caballé, respectively, then their

---

[21]Some existing work on increased assortative mating attributes it to an increased level of education among females. As above, one needs to be careful not to confuse this story with the effect of an increase in average educational attainment. Suppose, for instance, that the primary driver of assortative mating is wealth. Suppose also that there has been no change in the tendency for people to assortatively mate. If the daughters of the rich experience an increase in education that is larger than the daughters of the poor — something that is very plausible — then one might mistakenly conclude that assortative mating has increased, in the educational dimension, although in reality it has not. Our methodology does not suffer from this possible bias because our measures do not refer to the individual woman, *but to her family*. Education and ethnicity are imputed by the surname, not measured at an individual level.

complete surname is "Fernández Caballé". This is distinct from both "Caballé Fernández" and "Fernández Vila". Next, group each person together with those who share their complete surname, and eliminate anyone whose complete surname is unique in the population. This partition will be very similar to the actual sibling partition (which we cannot observe), with the similarity increasing in the *rarity* of the surnames. The reason is that it's very unlikely for two males who share the same rare surname to marry two females who share the same rare surname, thus generating children *who are not siblings* with the same complete surname. What is much more likely is that two individuals with the same *rare*, complete surname are in fact siblings.

We implement this as follows. Starting with the complete census population, we first eliminate all individuals with a unique complete-surname. Next, we form an increasingly large (and nested) set of subpopulations. The first is all those who share their complete surname with just one other person. The second adds to the first by including those who share their complete surname with two other people. We continue until we get to the sixth set, which includes the fifth and the remainder of the entire population. Finally, for each of these subpopulations we form the complete-surname partition using dummy variables and measure the ICS. Our model predicts that it should decrease as the subpopulation grows. That is, the first subpopulation consists of those with the rarest complete-surnames, and these people are almost surely siblings. Less so for the second subpopulation, and so on.

[Table 12 about here.]

Table 12(a) shows our results. Each column reports the R^2 of two regressions, one with complete-surname dummies, and another with "fake–complete–surname" dummies. The table also reports the associated ICS. Columns 1 through 6 represent the increasing set of subpopulations described above.

Table 12(a) is strongly supportive of our model. Both the R^2 and the ICS are much higher than in previous sections; surnames are more informative for the population of siblings. As predicted, the ICS declines — from 22.4% to 13.4% — as we increase the likelihood of spuriously grouping together individuals who are not siblings. Finally, as should be the case, the "fake–complete–surnames" R^2 are roughly the same as in previous sections (the small differences being attributable to different populations).

Table 12(b) repeats the analysis of Table 12(a), but using a much more ethnically homogeneous

overall population, those with the most Catalan surnames. The results, re-assuringly, are quite similar. We do this because Table 12(b) does not control for ethnicity using our *CatalanDegree* variable. To do so might pose multicollinearity problems because *CatalanDegree* and the complete-surname dummy would be based on a common surname.

These results are supportive of both our overall method and the approximate manner in which we form the siblings partition. We also ask if the sibling analysis is supportive of our 'dynamic results' in Section 9. We find that the same pattern arises. Table 13 shows the tables for young and old using complete surnames. Figures 13 and 14 show, respectively, the evolution over time for the entire population as well as the subpopulations with the more Catalan and less frequent surnames. The sibling-based ICS has increased over time.

[Table 13 about here.]

[Figure 13 about here.]

[Figure 14 about here.]

12 Conclusions

Our paper makes two contributions, one methodological and one applied. Methodologically, we develop a framework that shows how an untapped data source can shed light on a question that requires much data, but for which relatively little data exists. We show that a single cross-sectional census can reveal much about both the level and the change in intergenerational mobility. The key data objects are surnames, *markers* that provide intergenerational links where more explicit links are unavailable. Surnames define a *partition* of a population. Elements of this partition associated with *rare* surnames will be correlated with the partition that groups people according to familial linkages. A particular moment of these partitions — that which we label the *Informational Content of Surnames* (ICS) — connects the familial linkages with familial economic status and thus provides information on intergenerational mobility. This method yields measures of the degree of mobility at a point in time as well as its evolution across time. Our method also has the potential of comparing mobility across countries, as the model can be calibrated to accommodate differences in the surname distribution. We leave this to future research.

Our method would be of limited practical value in the presence of multi-country, intergenerational panel data. However the existence of such data is quite limited. The practical relevance of our method, therefore, depends on how much data we *do* have on the joint distribution of surnames and economic outcomes. Here, there is reason to be optimistic. Most countries compile censuses containing such data. We've shown that one can learn much from *one* census. Multiple census, both within and across countries, can obviously yield much more. *Comparability*, over time and across countries, can be handled. The US, for example, has a different surname distribution than Spain. The essence of our method — the idea that rare surnames connect people with familial linkages — is nevertheless unaffected.

Our practical contribution is to use our methodology to ask how and why intergenerational mobility has *changed* over time. We study Catalonia, a large region of Spain. Using the 2001 census we show that the explanatory power of surnames — the ICS — has increased. Part of this is due to the increased explanatory power of ethnicity. But there is more going on. There is a component of the ICS that is *unrelated* to ethnicity and the impact of this component has also increased. This is true among very ethnically-homogeneous individuals, among siblings, and among people with infrequent surnames. Our model, alongside an extensive set of controls and robustness checks, associates this increase in the ICS with a decrease in intergenerational mobility. Our model and data also offer one possible explanation. Assortative mating along the ethnic dimension appears to have increased in tandem with the decrease in mobility.

To wrap-up, we offer some historical context. In Spain and Catalonia, the different generations of the 20th century witnessed large-scale *increases* in both the level of publicly-provided education and the level of educational attainment. Nevertheless, we've found that educational mobility has *decreased*. That is, the importance of family-specific characteristics for educational outcomes has *increased*. Is their a logical contradiction here? If one looks around and sees that almost everyone's educational attainment exceeds that of their parents, does this mean that the importance of inheritance and familial linkages must have diminished? The answer is no. Such logic confuses aggregate growth — an increase in the *mean* of the distribution — with mobility, which is all about movement *within* the distribution. It is at the heart of the common misperception that to do better than one's parents means to have beaten the odds and done better than expected. This can generate an upward bias in our perception of intergenerational mobility in growing economies. It is an illusion. It is just growth. Mobility works along its own path. It is defined only in relative

terms. To measure mobility, it is not enough to compare my welfare with that of my parents. I must also consider the children of other parents, parents that were both richer and poorer than my own. Today's generation may well live better than yesterday's, while at the same owing a greater thanks to their parents for their place in the cross-sectional distribution.

References

Aaronson, D. and B. Mazumder (2008). Intergenerational economic mobility in the U.S.: 1940 to 2000. *Journal of Human Resources 43*(1), 139–172.

Angelucci, M., G. De Giorgi, M. Rangel, and I. Rasul (2010). Family networks and school enrolment: Evidence from a randomized social experiment. *Journal of Public Economics 94*(3-4), 197–221.

Aspachs-Bracons, O., I. Clots-Figueras, J. Costa-Font, and P. Masella (2008). Compulsory language educational policies and identity formation. *Journal of the European Economic Association 6*(2-3), 434–444.

Bagüés, M. F. (2005). ¿Qué determina el éxito en unas oposiciones? Fedea, Documento de Trabajo 2005-01.

Becker, G. S. (1967). Human capital and the personal distribution of income: An analytical approach. Woytinsky Lecture, Institute of Public Administration, University of Michigan.

Becker, G. S. (1973). A theory of marriage: Part i. *Journal of Political Economy 81*, 813–846.

Becker, G. S. and N. Tomes (1979). An equilibrium theory of the distribution of income and intergenerational mobility. *Journal of Political Economy 87*(6), 1153–1189.

Becker, G. S. and N. Tomes (1986). Human capital and the rise and fall of families. *Journal of Labor Economics 4*(3 part 2), S1–S39.

Behrmana, J. and P. Taubman (1985). Intergenerational earnings mobility in the United States: Some estimates and a rest of becker's intergenerational endowments model. *Review of Economic and Statistics 67*, 144–51.

Behrmana, J. and P. Taubman (1990). The inergenerational correlation between children's adults earnings and their parents' income: Results from the Michigan panel survey of income dynamics. *The Review of Income and Wealth 36*(2), 115–127.

Bertrand, M. and S. Mullainathan (2004). Are Emily and Greg more employable than Lakisha and Jamal? a field experiment on labor market discrimination. *American Economic Review 94*(4), 991–1013.

Björklund, A., T. Eriksson, M. Jäntti, O. Raaum, and E. Österbacka (2002). Brother correlations in earnings in Denmark, Finland, Norway, and Sweden compared to the United States. *Journal of Population Economics 15*(4), 757–772.

Björklund, A. and M. Jäntti (1997). Intergenerational income mobility in Sweden compared to the United States. *American Economic Review 87*(5), 1009–1018.

Black, S. E. and P. J. Devereux (2011). *Recent Developments in Intergenerational Mobility.* in Orley C. Ashenfelter and David Card (eds.), Handbook of Labor Economics, Volume 4B, Amsterdam: North-Holland, pp. 1487-1541.

Blanden, J., A. Goodman, P. Gregg, and S. Machin (2004). *Changes in Intergenerational Mobility in Britain.* in Miles Corak (ed.), Generational Income Inequality, Cambridge University Press, pp.122-146.

Cabré, A. (2004). La aportación de los 'otros' catalanes. El País (Edición Barcelona), 06/09/04.

Chadwick, L. and G. Solon (2002). Intergenerational income mobility among daughters. *American Economic Review 92*(1), 335– 44.

Checchi, D., A. Ichino, and A. Rustichini (1999). More equal but less mobile? education financing and intergenerational mobility in Italy and in the U.S. *Journal of Public Economics 74*(3), 351.93.

Clark, G. (2010). Regression to mediocrity? surnames and social mobility in England, 1200-2009. mimeo.

Collado, M. D., I. Ortuño-Ortín, and A. Romeo (2006). Vertical transmission of consumption behavior and the distribution of surnames. mimeo.

Comi, S. (2003). Intergenerational mobility in Europe: evidence from ECHP. mimeo.

Couch, K. A. and T. A. Dunn (1997). Intergenerational correlations in labor market status: A comparison of the United States and Germany. *Journal of Human Resources 32*(1), 210–32.

Dahan, M. and A. Gaviria (2001). Sibling correlations and intergenerational mobility in Latin America. *Economic Development and Cultural Change, University of Chicago Press 49*(3), 537–54.

Darwin, G. H. (1875). Marriages between first cousins in England and their effects. *Journal of the Statistical Society 38*, 153–184.

Dearden, L., S. Machin, and H. Reed (1997). Intergenerational mobility in Britain. *Economic Journal 107*, 47–64.

Duncan, O. D., D. Featherman, and B. Duncan (1972). *Sociological Background and Achievement.* New York: Seminar Press.

Dunn, C. (2007). The intergenerational transmission of lifetime earnings: Evidence from Brazil. *The B.E. Journal of Economic Analysis & Policy 7*,(Iss. 2 (Contributions)), Article 2.

Ermisch, J., M. Francesconi, and T. Siedler (2006). Intergenerational mobility and marital sorting. *Economic Journal 116*, 659–679.

Ferreira, S. G. and F. A. Veloso (2006). Intergenerational mobility of wages in Brazil. *Brazilian Review of Econometrics 26*(2), 181–211.

Fertig, A. R. (2004). "trends in intergenerational earnings mobility in the U.S.". *Journal of Income Distribution 12*, 108–130.

Fryer, R. and S. Levitt (2004). The causes and consequences of distinctively black names. *Quarterly Journal of Economics 119*(3), 767–805.

Gavilán, A. (2011). Wage inequality, segregation by skill and the price of capital in an assignment model. *European Economic Review, forthcoming.*

Grawe, N. D. (2004). *Intergenerational mobility for whom? The experience of high- and low-earnings son in international perspective.* in Miles Corak (ed.), Generational Income Inequality, Cambridge University Press, pp.58-89.

Haider, S. and G. Solon (2006). Life-cycle variation in the association between current and lifetime earnings. *The American Economic Review 96*(4), 1308–1320.

Hertz, T. (2007). Trends in the intergenerational elasticity of family income in the United States. *Industrial Relations 46 (1)*, 22–50.

Hertz, T. N. (2001). Education, inequality andeconomic mobility in South Africa. Ph.D. thesis, University of Massachusetts.

Holmlund, H. (2006). Intergenerational mobility and assortative mating: Effects of an educational reform. working paper 4/2006, Swedish Institute for Social Research, Stockholm University.

Kremer, M. and E. Maskin (1995). Wage inequality and segregation by skill. NBER Working Paper num. 5718.

Lam, D. and R. F. Schoeni (1993). Effects of family background on earnings and return to schooling: evidence from brazil. *Journal of Political Economy 101*(4), 710–40.

Lasker, G. W. (1985). *Surnames and genetic structure.* Cambridge: Cambridge University Press.

Lee, C.-I. and G. Solon (2006). Trends in intergenerational income mobility. NBER WP 12007.

Leigh, A. (2007). Intergenerational mobility in Australia. *The B.E. Journal of Economic Analysis & Policy 7,*(Iss. 2 (Contributions)), Article 6.

Levine, D. I. and B. Mazumder (2007). The growing importance of family: Evidence from brothers' earnings. *Industrial Relations 46 (1)*, 7–21.

Levitt, S. and S. J. Dubner (2005). *Freakonomics: A Rogue Economist Explores the Hidden Side of Everything.* William Morrow/HarperCollins.

Lillard, L. A. and M. R. Kilburn (1995). Intergenerational earnings links: Sons and daughters. Papers 95-17, RAND - Labor and Population Program.

Long, J. and J. Ferrie (2011). Intergenerational occupational mobility in Britain and the U.S. since 1850. *American Economic Review, forthcoming.*

Manrubia, S. C. and D. H. Zanette (2002). At the boundary between biological and cultural evolution: The origin of surname distributions. *Journal of Theoretical Biology 261*(4), 461–477.

Marimon, R. and F. Zilibotti (1999). Unemployment vs. mismatch of talents: Reconsidering unemployment benefits. *Economic Journal 109*(455), 266–291.

Mayer, S. E. and L. M. Lopoo (2005). Has the intergenerational transmission of economic status changed? *Journal of Human Resources 40*(1), 169–85.

Miley, T. J. (2004). The politics of language and nation: The case of the Catalans in contemporary Spain. Ph.D. thesis, Department of Political Science at Yale University.

Ng, I. (2007). Intergenerational income mobility of young singaporeans. *The B.E. Journal of Economic Analysis & Policy 7,*(Iss. 2 (Topics)), Article 3.

Olivetti, C. and D. Paserman (2011). In the name of the father: Marriage and intergenerational mobility in the United States, 1850-1930. mimeo.

Osterbacka, E. (2001). Family background and economic status in Finland. *Scandinavian Journal of Economics 103*(3), 467– 84.

Osterberg, T. (2000). Intergenerational income mobility in Sweden: What do tax data show? *Review of Income and Wealth. 46*(4), 421–36.

Page, M. E. and G. Solon (2003). Correlations between brothers and neighboring boys in their adult earnings: The importance of being urban. *Journal of Labor Economics 21*, 831–55.

Shimer, R. and L. Smith (2000). Assortative matching and search. *Econometrica 2*(68), 343–369.

Solon, G. (1992). Intergenerational income mobility in the United States. *American Economic Review 82*(3), 393–408.

Solon, G. (1999). *Intergenerational Mobility in the Labor Market.* in Orley C. Ashenfelter and David Card (eds.), Handbook of Labor Economics, Volume 4B, Amsterdam: North-Holland, pp. 1761-1800.

Solon, G. (2002). Cross-country differences in intergenerational earnings mobility. *Journal of Economic Perspectives 16*(3), 59– 66.

Solon, G., M. Corcoran, Roger, and L. Deborah (1991). A longitudinal analysis of siblings correlations in economic status. *Journal of Human Resources 26*, 509–34.

Wiegand, J. (1997). Intergenerational earnings mobility in Germany. mimeo.

A   Appendix: Proofs of Model's Properties

Proof of Property 1

Consider some date $t-1$ and a surname $s \in \Omega$ such that $F_{t-1}(s) > 0$. Define $Q_{t-1}(s)$ as the number of individuals with surname s so that $Q_{t-1}(s) = N_{t-1}F_{t-1}(s)$. Note that, conditional on $Q_{t-1}(s)$, $Q_t(s)$ is a binomial random variable with support $[0, m\, Q_{t-1}(s)]$ and distribution (suppressing the '(s)' notation)

$$\text{Prob}(Q_t = km) = \left(\begin{array}{c} Q_{t-1} \\ k \end{array} \right) q^k (1-q)^{Q_{t-1}-k} \ . \tag{12}$$

The conditional mean and variance of Q_t are $Q_{t-1}mq$ and $Q_{t-1}m^2q(1-q)$, respectively. Therefore,

$$Q_t \quad = \quad Q_{t-1}mq + w_t \ , \tag{13}$$

where w_t is defined as the innovation, $w_t \equiv Q_t - E_{t-1}Q_t$. If $mq = 1$ then Q_t follows a driftless random walk.[22]

Proof of Property 3

Fix some date t. Partition the population into *families*: groups of individuals who share the same lineage (which is possible because of asexual reproduction). Suppose, to begin with, that every family's lineage dates back k periods and that no two families share the same surname. Then the cross-sectional mean and variance of income for each family are, respectively,

$$E(y_{ist} \,|\, s) \quad = \quad \rho^k y_{s,t-k} \tag{14}$$

$$Var(y_{ist} \,|\, s) \quad = \quad V_\varepsilon \sum_{l=0}^{k-1} \rho^{2l} \tag{15}$$

where $y_{s,t-k}$ is the income of the patriarch of the family with surname s. Now recognize that the society-wide cross-sectional variance can be decomposed into the average within-family variance and across-family variance in the conditional mean:

$$Var(y_{ist}) \quad = \quad E\big(Var(y_{ist} \,|\, s) \big) + Var \big(E(y_{ist} \,|\, s) \big) \ . \tag{16}$$

---

[22]This is related to the "branching process" literature. It was started by Francis Galton in 1873. He posed the following problem (our model with zero mutation).

> *Problem 4001:* A large nation, of whom we will only concern ourselves with the adult males, N in number, and who each bear separate surnames, colonise a district. Their law of population is such that, in each generation, a_0 percent of the adult makes have no make children who reach adult life; a_1 have one such male child; a_2 have two; and so on up to a_5 who have five.
>
> Find (1) what proportion of the surnames will have become extinct after r generations; and (2) how many instances there will be of the same surname being held by m persons.

The answer was finally figured out using martingale methods, but not until 1950! It's kind of complicated, but the upshot is that, with strictly positive population growth a fraction q of all surnames with vanish with probability 1 and a fraction $(1-q)$ will persist forever (U.S. data on q suggests about 0.8). The distribution for the surviving names is exponential. This is from "Branching processes since 1873," by David Kendall. Google this title and you'll find it right away.

The ICS from equation (5) is proportional to the second term on the right which, according to expression (14) is monotonically increasing in ρ. This proves Property 3 for the case identical lineage horizons and unique within-family surnames.

Consider next the general case of lineage horizons that vary across families. Suppose that family j all derive from a patriarch who lived k_j periods before date t. Then equations (14) and (15) remain valid for each k_j and equation 16 takes the form of family-size weighted means and variances. Nevertheless, holding fixed the structure of the population, the second term on the right of equation (16) remains a monotonically increasing function of ρ.

Finally, relax the assumption that surnames and families are uniquely associated. If family j_1 and family j_2 share the same surname, s, then $E(y_{ist}\,;\,s)$ is a family-size-weighted average of the incomes of all of the members of the two families. Such averaging will, of course, decrease the cross-sectional variance in the conditional means, $Var(E(y_{ist}\,|\,s))$, thereby decreasing the ICS. However, holding fixed the population structure, it remains the case that this conditional variance, and the ICS, are increasing in ρ.

Proof of Property 4

Here, we demonstrate that our assortative mating model from Section 5.1 has a unique stationary distribution and derive expressions for the models variances and correlations in terms of its structural parameters.

Recall that the male and female children in the ith household with surname s at date t have income described by

$$y_{ist}^m = rz_{ip,t-1} + e_{ist}^m \quad ; \quad y_{ist}^f = rz_{ip,t-1} + e_{ist}^f \quad , \tag{17}$$

where $z_{ip,t-1}$ is the average income of these children's parents, who formed this household at date $t-1$, r is the *household* inheritance parameter and the innovations e are *i.i.d.* $N(0, V_e)$. We now suppress the i and s notation (they are not needed here). Mating is described by

$$y_{pt}^f = \lambda y_{pt}^m + u_{pt} \quad ; \quad u_{pt} \sim N(0, V_u) \quad . \tag{18}$$

First, we guess that there exists a stationary distribution for z that has the form $N(0, V_z)$. If so, then parental income at date $t+1$ — formed from the date t mating rule (18) — satisfies

$$
\begin{aligned}
z_{pt} &= \left(y_{pt}^m + y_{pt}^f\right)/2 \\
&= \left((1+\lambda)y_{pt}^m + u_{pt}\right)/2 \quad ,
\end{aligned}
$$

where the first equation is just the definition of average parental income and the second applies the mating rule (18). Applying the inheritance process (17), we get

$$2z_{pt} = \left(rz_{p,t-1} + e_t^m\right)\left(1+\lambda\right) + u_{pt} \quad .$$

The variance of the distribution of z, then (if it exists), results from taking the unconditional variance of both sides and imposing stationarity:

$$V_z = \frac{(1+\lambda)^2 V_e + V_u}{4 - (1+\lambda)^2 r^2} \quad . \tag{19}$$

This gives V_z as a function of the structural parameters λ, V_e and r, and the variance of mating noise, V_u,

which is uniquely determined below.

Next, note that a stationary distribution for z implies that the income of male and female children have the same distribution (*i.e.*, by inspection of Equation (17)). Thus, we can write $y_{ist}^m \sim N(0, V_y)$ and $y_{ist}^f \sim N(0, V_y)$, for some variance, V_y, to be uniquely determined below. Given this, the mating rule, Equation (18), imposes that

$$V_u = (1 - \lambda^2)V_y \ . \tag{20}$$

This guarantees that the distribution of female income implied by the mating rule coincides with that implied by the inheritance process.

Next, consider the income of males at date $t + 1$.

$$
\begin{aligned}
y_{t+1}^m &= rz_{pt} + e_{t+1}^m \\
&= r\big((1 + \lambda)y_{pt}^m + u_{pt}\big)/2 + e_{t+1}^m \\
&= \frac{r(1 + \lambda)}{2}y_{pt}^m + ru_{pt}/2 + e_{t+1}^m.
\end{aligned}
\tag{21}
$$

Since $r < 1$ and $\lambda < 1$, then $r(1 + \lambda)/2 < 1$. Given the independence assumptions on u and e, and given that fertility is deterministic (with each male bearing one male offspring), then Equation (21) gives the income of a male as stationary Gaussian first-order autoregressive function of the income of his father. Its unconditional distribution is

$$y_t^m \sim N\Big(0, \frac{r^2 V_u/4 + V_e}{1 - r^2(1 + \lambda)^2/4}\Big) \ .$$

By a cross-sectional law-of-large numbers, this also gives the stationary cross-sectional distribution of male income.

All that remains is to solve for V_y as a function of the model's structural parameters. Using this last expression,

$$
\begin{aligned}
V_y &= \frac{r^2 V_u/4 + V_e}{1 - r^2(1 + \lambda)^2/4} \\
&= \frac{r^2 V_y(1 - \lambda^2)/4 + V_e}{1 - r^2(1 + \lambda)^2/4} \\
&= \frac{V_e}{\lambda(1 + \lambda)} \ ,
\end{aligned}
$$

where the second equation follows from Equation (20) and the third follows from solving for V_y and rearranging. Substituting the result into Equation (19) yields:

$$V_z = \frac{\lambda(1 + \lambda)^2 V_e + 2(1 - \lambda)V_e}{\lambda(4 - (1 + \lambda)^2 r^2)}$$

This implies that there does indeed exist a stationary distribution for average parental income, z, that is consistent with the inheritance and mating rules, (17) and (18). The variance of the *male* inheritance shock, w_{ist}^m from Equation (10), is

$$V_w = V_e\Big(1 + \frac{r^2(1 - \lambda)}{4\lambda}\Big).$$

50

(a) Time series of average number of agents per surname

(b) Times series of Gini of surname distribution

(c) Ordered frequency distribution, $G(k)$


(d) Lorenz curve of surname distribution
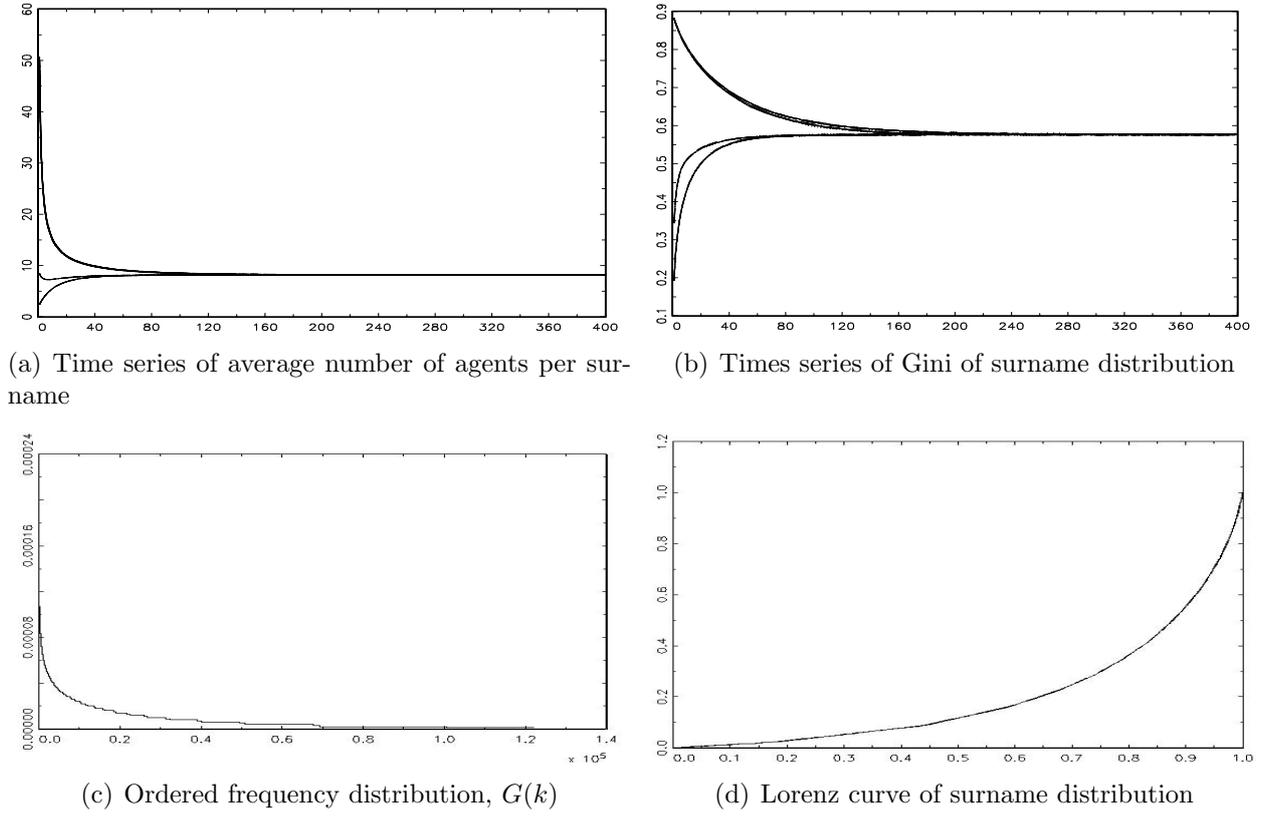
Figure 1: Time series of the average number of agents per surname & Gini coefficient, $G(k)$ and Lorenz curve, for different values of ρ.

Notes: Model Simumations with Baseline Parameter Values: N_0=1000000; V_ε=1.000; μ=0.0200; q=0.50; m=2; $\rho \in [0.05, 0.95]$. Different initial conditions: number of surnames= 10, 1000, 100000 and 1000000.

(a) Average R_L^2 (solid line) and R_F^2 (dotted line) against ρ
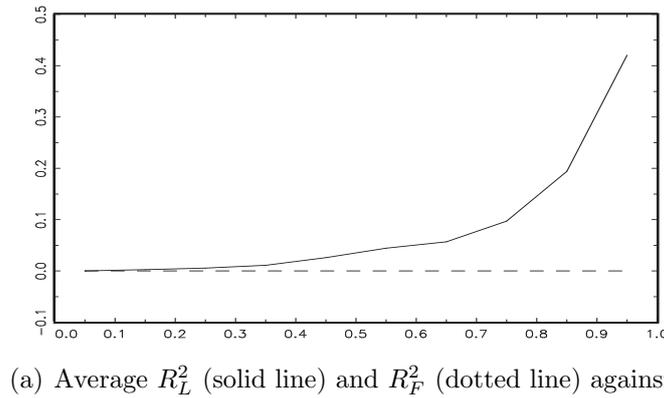
Figure 2: Surnames are informative, and their informational content increases with the degree of inheritance that there is in society.

Notes: Model Simumations with Baseline Parameter Values: as in Figure 1.
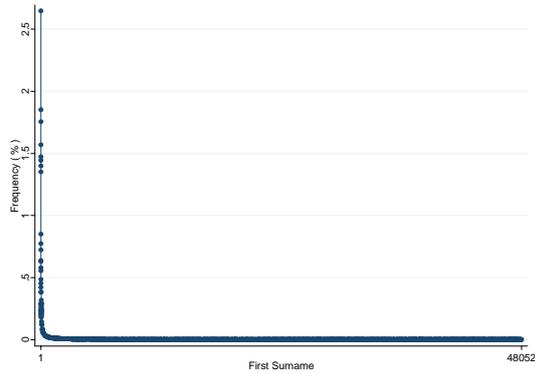
(a) Distribution of First Surname
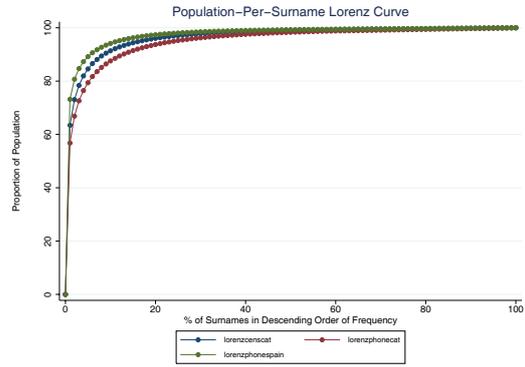
(b) Lorenz Curve of the Surname Distribution in Catalonia and Spain

Figure 3: Distribution of the first surname in Catalonia and lorentz curves for Spain and Catalonia

For 3(a): Source: 2001 Catalan Census. Sample: Spanish citizens living in Catalonia aged 25 and above, all surnames.
For 3(b): Source: 2004 Spanish Phone Book & 2001 Catalan Census. Sample: All phones with first & second surnames not missing. Population Percentage per Surname (1% Steps).
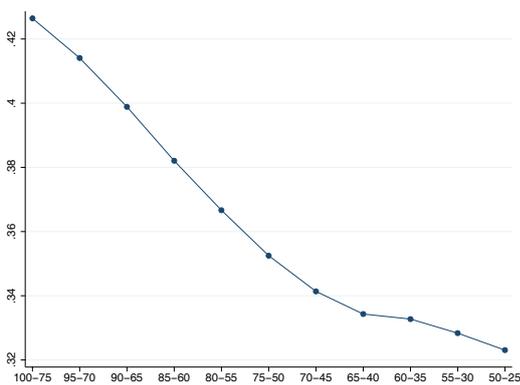

(a) Average of *CatalanDegree*
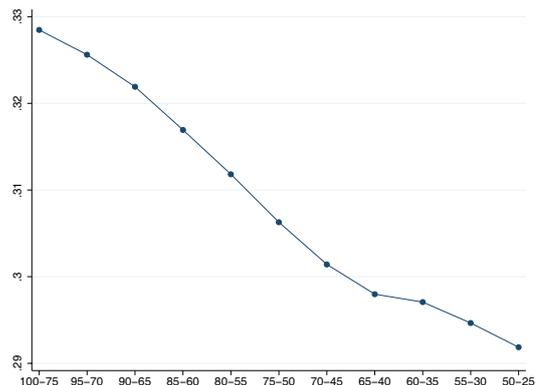
(b) Standard Deviation of *CatalanDegree*

Figure 4: Evolution of *CatalanDegree* over moving windows of cohorts

Source: 2001 Catalan Census. Samples: Overlaping age-cohorts (as described in footnote 14) of complete population.
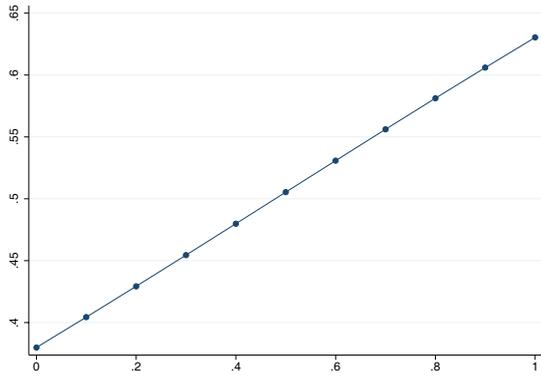
(a) Probability of full knowledge of Catalan language
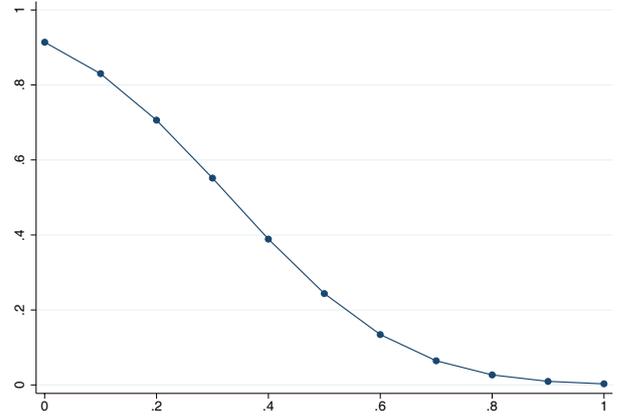
(b) Probability of being immigrant in Catalonia

Figure 5: Probabilities of knowledge of Catalan language and being an immigrant in Catalonia, as a function of *CatalanDegree*.

Notes and Samples: as in Table 3. For figure 5(a), reference individual is a male, aged 50-55, born in the county of Barcelona. For figure 5(b), reference individual is a male, aged 60-65.
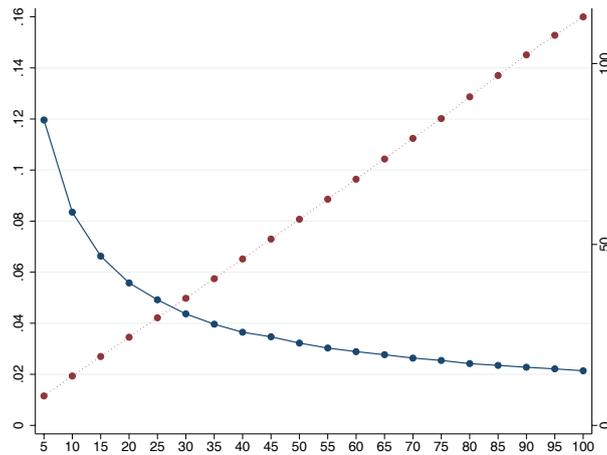

Figure 6: ICS is larger for less frequent surnames

Notes: ICS (solid line) and individuals per surname (dotted line). Regressions as in table 4 (columns 3 and 4) by percentiles, where percentile x corresponds to the $x\%$ least frequent surnames. Source: 2001 Catalan Census. Samples: $x\%$ least frequent surnames of complete population.
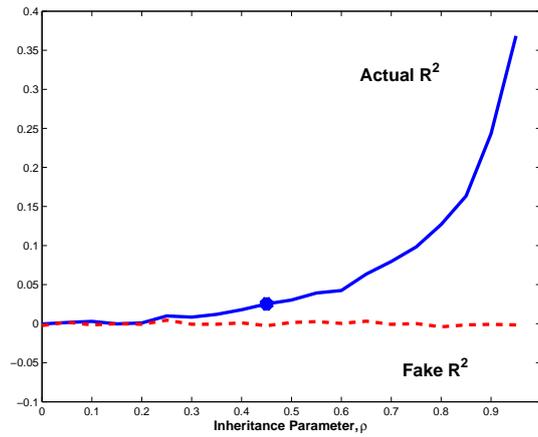
Figure 7: Calibration of Inheritance Parameter

Notes: Based on simulated data as described in Section 8. The solid blue (dashed red) line plots the R^2 from the actual (fake) dummy variable regression described by Equation (3) (Equation (4)) in Section 3. The big dot represents the calibrated value of $\rho = 0.47$ associated with an ICS of 2.13%.


(a) Average

(b) Standard Deviation
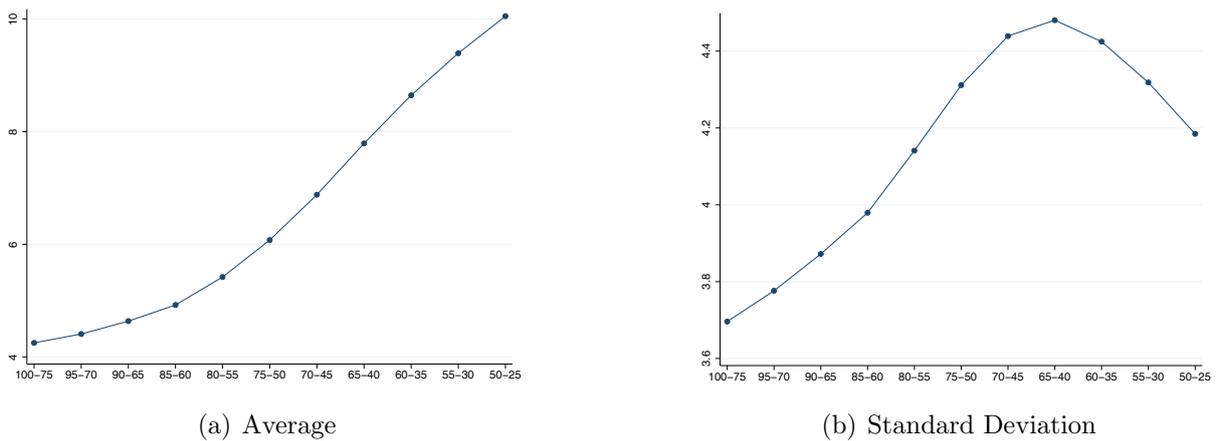
Figure 8: Evolution of years of education over moving windows of cohorts

Source: 2001 Catalan Census. Samples: Overlaping age-cohorts (as described in footnote 14) of complete population.
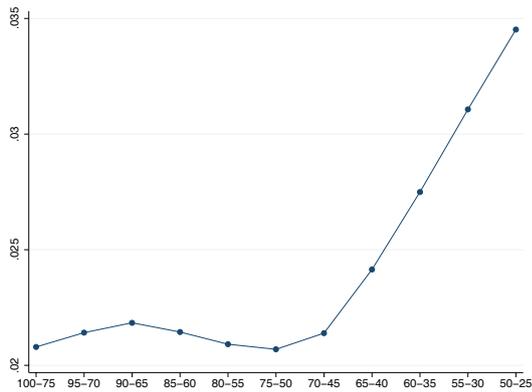
(a) Evolution of ICS
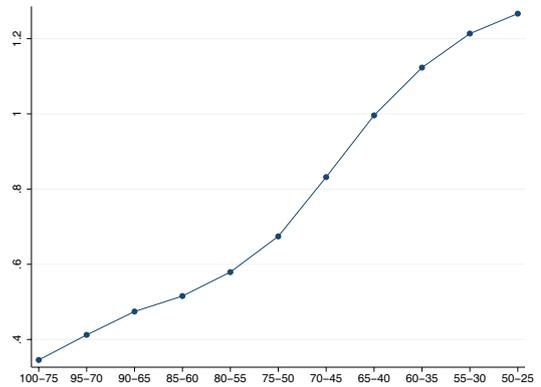
(b) Evolution of parameter of *CatalanDegree*

Figure 9: Evolution of ICS and parameter of *CatalanDegree* over moving windows of cohorts

Notes: Regressions as in table 4 (columns 3 and 4). Source: 2001 Catalan Census. Samples: Overlapping age-cohorts (as described in footnote 14) of complete population.
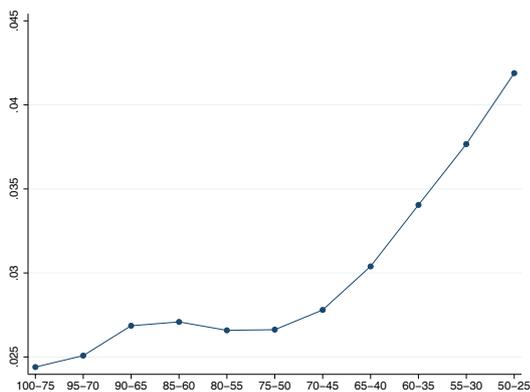
(a) 50% Most Catalan Surnames
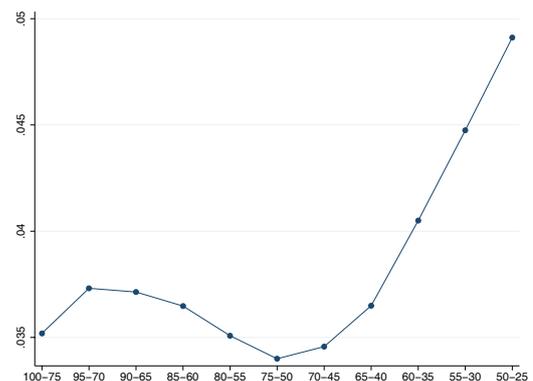
(b) 50% Least Frequent Surnames

Figure 10: Evolution of ICS over moving windows of cohorts, subpopulations.

Notes: Regressions as in table 4 (columns 3 and 4). Source: 2001 Catalan Census. Samples: Overlapping age-cohorts (as described in footnote 14) of subpopulations of complete population.
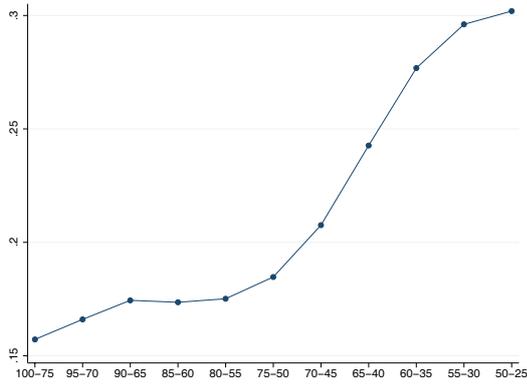
(a) AM in Education
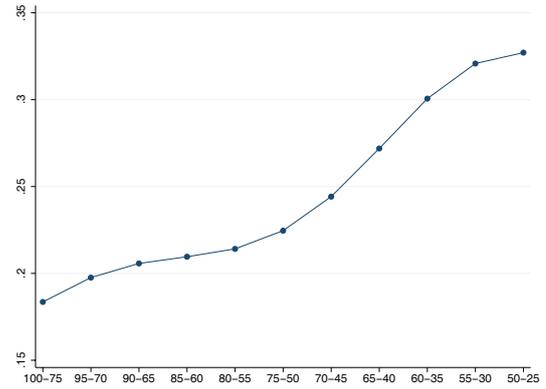
(b) AM in *CatalanDegree*

Figure 11: Evolution of Assortative Mating in Education & *CatalanDegree* over moving windows of cohorts.

Notes: Regressions as in table 11. Source: 2001 Catalan Census. Samples: Overlapping age-cohorts (as described in footnote 14) of subpopulations of sample in table 11.
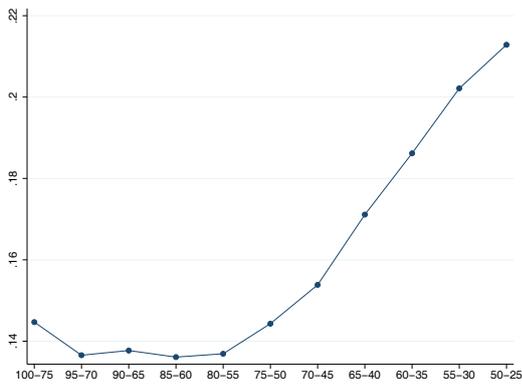
(a) 50% Most Catalan Surnames
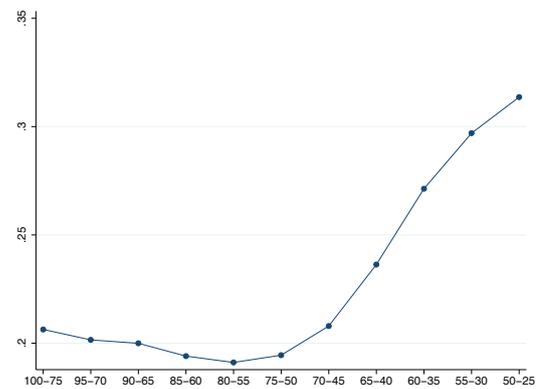
(b) 50% Least Frequent Surnames

Figure 12: Evolution of Assortative Mating in Education over moving windows of cohorts, sub-populations.

Notes: Regressions as in table 11(a). Source: 2001 Catalan Census. Samples: Overlapping age-cohorts (as described in footnote 14) of subpopulations of sample in table 11.
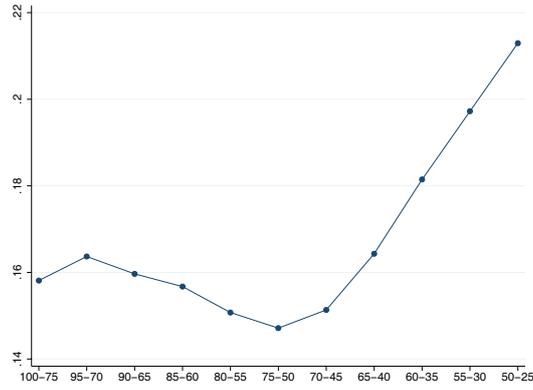
Figure 13: Evolution of ICS over moving windows of cohorts. Complete Surnames ("Siblings")

Notes: Regressions as in table 12(a). Source: 2001 Catalan Census; Sample: Overlaping age-cohorts (as described in footnote 14) of sample in table 12(a), column 6.
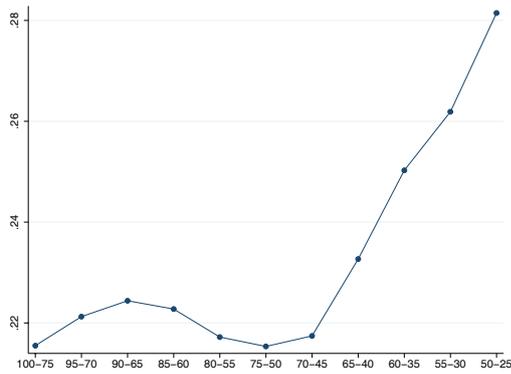

(a) 50% Most Catalan Surnames
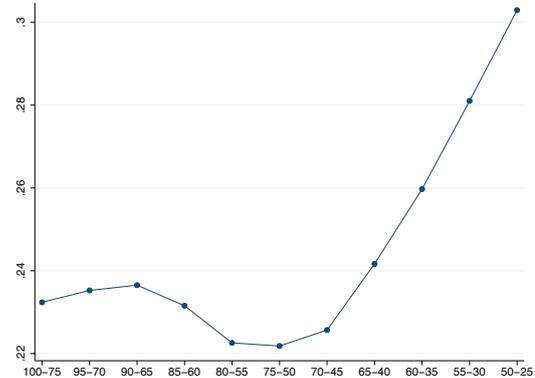
(b) 50% Least Frequent Surnames

Figure 14: Evolution of ICS over moving windows of cohorts. Complete Surnames ("Siblings"). Subpopulations.

Notes: Regressions as in table 12(a). Source: 2001 Catalan Census; Sample: Overlapping age-cohorts (as described in footnote 14) of subpopulations of sample in table 12(a), column 6.

Table 1: Surnames Distribution: Gini Index and People per Surname in Catalonia and Spain

	Spain (PhoneBook)	Catalonia (Census)	Catalonia (PhoneBook)
Number of People	11,397,116	6,123,909	2,073,219
Number of Surnames	155,782	91,568	61,396
People per Surname	73.161	66.878	33.768
Gini Index	0.9485	0.9304	0.9028

Source: 2004 Spanish Phone Book & 2001 Catalan Census. Sample: All individuals/phones with first & second surnames not missing.

Table 2: *CatalanDegree* Summary Statistics

	All residents in Catalonia	Born in Catalonia before 1950	Born anywhere in Spain before 1950	after 1950
	(1)	(2)	(3)	(4)
Mean CatalanDegreeSurname2	0.344	0.5672	0.367	0.322
Standard deviation	(0.302)	(0.3241)	(0.312)	(0.292)
Share with CatalanDegreeSurname2>0.16	0.568	0.8365	0.596	0.542

Source: 2001 Catalan Census. Samples: Complete population and subpopulations.

Table 3: *CatalanDegree* & Probabilities of knowledge of Catalan language and of being an immigrant

(a) Probability of knowledge of Catalan language

LHS: Knowledge of Catalan	(1)	(2)
CatalanDegreeSurname2		0.639 $_{(.003)}$
Log likelihood	-2248320.8	-2219774.9
Pseudo R^2	0.2387	0.2483

(b) Probability of being immigrant

LHS: Immigrant	(1)	(2)
CatalanDegreeSurname2		-4.121 $_{(.006)}$
Log likelihood	-1418949.8	-903746.31
Pseudo R^2	0.0052	0.3664

Notes: Probit Estimates. All regressions include gender and age dummies. Regressions in table 3(a) also include place of birth dummies. Standard errors in parenthesis. Source: 2001 Catalan Census. For 3(a): Sample: Complete population. The LHS variable *Knowledge of Catalan* takes value 1 for individuals who understand, can speak, can read and can write the Catalan language and zero otherwise. Number of observations: 4,293,173. For 3(b): Sample: Individuals born before 1950 of complete population. The LHS variable *Immigrant* takes value 1 for individuals who were not born in Catalonia and zero otherwise. Number of observations: 2,057,831.

Table 4: ICS. Spanish citizens living in Catalonia.

LHS: years of education	(1)	(2)	(3)	(4)	(5)	(6)
CatalanDegreeSurname2		$1.692_{(0.007)}$	$1.017_{(0.008)}$	$1.692_{(0.007)}$		
Surname Dummies			Yes		Yes	
Fake Surnames Dummies				Yes		Yes
Adjusted R-squared	0.3363	0.3440	0.3653	0.3440	0.3629	0.3363
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.384	0.000	0.332

Notes: All regressions include gender, age and place of birth dummies. Fake-surnames have the same distribution as Surnames and are allocated randomly. (*) F-test if Surname dummies are jointly significant. Standard errors in parenthesis. Source: 2001 Catalan Census. Sample: Spanish citizens living in Catalonia aged 25 and above, with frequency of first surname larger than one. Number of observations: 4,293,173. Number of surnames: 38,024.

Table 5: ICS. Subpopulations.

(a) Born in Catalonia.

LHS: years of education	(1)	(2)	(3)	(4)	(5)	(6)
CatalanDegreeSurname2		$1.703_{(0.008)}$	$0.994_{(0.009)}$	$1.703_{(0.008)}$		
Surname Dummies			Yes		Yes	
Fake Surnames Dummies				Yes		Yes
Ajusted R-squared	0.2256	0.2379	0.2697	0.2378	0.2661	0.2255
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.862	0.000	0.855

(b) Born in Catalonia before 1950. ("Old")

	(1)	(2)	(3)	(4)	(5)	(6)
CatalanDegreeSurname2		$0.952_{(0.013)}$	$0.575_{(0.014)}$	$0.956_{(0.013)}$		
Surname Dummies			Yes		Yes	
Fake Surnames Dummies				Yes		Yes
Adjusted R-squared	0.1578	0.1622	0.1947	0.1621	0.1932	0.1577
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.709	0.000	0.772

(c) 50% Most Catalan Surnames.

	(1)	(2)	(3)	(4)	(5)	(6)
CatalanDegreeSurname2		$0.972_{(0.010)}$	$0.808_{(0.010)}$	$0.971_{(0.010)}$		
Surname Dummies			Yes		Yes	
Fake Surnames Dummies				Yes		Yes
Adjusted R-squared	0.3213	0.3246	0.3488	0.3245	0.3467	0.3212
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.947	0.000	0.939

Notes: Regressions as in table 4. Source: 2001 Catalan Census. For 5(a): Sample: Individuals born in Catalonia of complete population. Number of observations: 2,721,917. Number of surnames: 31,987. For 5(b): Sample: Individuals born in Catalonia before 1950 of complete population. Number of observations: 1,017,123. Number of surnames: 21,602. For 5(c): Sample: Individuals with 50% Most Catalan Surnames of complete population. Number of observations: 2,146,587. Number of surnames: 28,862.

Table 6: ICS. 50% Least Frequent Surnames.

LHS: years of education	(1)	(2)	(3)	(4)	(5)	(6)
CatalanDegreeSurname2		$1.459_{(0.010)}$	$0.815_{(0.011)}$	$1.459_{(0.010)}$		
Surname Dummies			Yes		Yes	
Fake Surnames Dummies				Yes		Yes
Adjusted R-squared	0.3316	0.3379	0.3701	0.3379	0.3684	0.3316
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.319	0.000	0.358

Notes: Regressions as in table 4. Source: 2001 Catalan Census. Sample: Individuals with 50% Least Frequent Surnames (by Surname1) of complete population. Number of observations: 2,147,479. Number of surnames: 37,685.

Table 7: ICS. "Invented" Catalonias.

LHS: years of education	(1)	(2)	(3)
CatalanDegreeSurname2	Yes	Yes	Yes
Adjusted $R^2_{surnames}$	0.3653	0.3661	0.3646
Adjusted R^2_{fake}	0.3440	0.3429	0.3452
ICS	0.0213	0.0232	0.0194
Sample	All surnames	"First letters"	"Last letters"
Observations	4,293,173	2,187,084	2,106,089

Notes: Regressions as in table 4 (columns 3 and 4). Source: 2001 Catalan Census. Samples: Column (1): complete population; Columns (2) and (3) are the first and second half respectively of the complete population alphabetically ordered.

Table 8: ICS over cohorts.

(a) Born before 1950 ("Old")

LHS: years of education	(1)	(2)	(3)	(4)	(5)	(6)
CatalanDegreeSurname2		0.896	0.594	0.897		
Surname Dummies			Yes		Yes	
Fake Surnames Dummies				Yes		Yes
Adjusted R-squared	0.2313	0.2335	0.2533	0.2335	0.2524	0.2313
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.679	0.000	0.688

(b) Born after 1950 ("Young")

CatalanDegreeSurname2		2.143	1.271	2.145		
Surname Dummies			Yes		Yes	
Fake Surnames Dummies				Yes		Yes
Adjusted R-squared	0.1002	0.1183	0.1534	0.1184	0.1481	0.1002
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.260	0.000	0.421

Notes: Regressions as in table 4. Source: 2001 Catalan Census.
For 8(a): Sample: Individuals born before 1950 of complete population. Number of observations: 2,052,725. Number of surnames: 31,237.   For 8(b): Sample: Individuals born after 1950 of complete population. Number of observations: 2,232,102. Number of surnames: 31,847.

Table 9: ICS over cohorts. 50% Most Catalan Surnames.

(a) Born before 1950 ("Old").

LHS: years of education	(1)	(2)	(3)	(4)	(5)	(6)
CatalanDegreeSurname2		$0.634_{(0.014)}$	$0.432_{(0.014)}$	$0.631_{(0.014)}$		
Surname Dummies			Yes		Yes	
Fake Surnames Dummies				Yes		Yes
Adjusted R-squared	0.2122	0.2137	0.2397	0.2137	0.2390	0.2122
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.455	0.000	0.417

(b) Born after 1950 ("Young").

	(1)	(2)	(3)	(4)	(5)	(6)
CatalanDegreeSurname2		$1.790_{(0.013)}$	$1.075_{(0.013)}$	$1.790_{(0.013)}$		
Surname Dummies			Yes		Yes	
Fake Surnames Dummies				Yes		Yes
Adjusted R-squared	0.0858	0.1018	0.1444	0.1017	0.1394	0.0857
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.576	0.000	0.569

Notes: Regressions as in table 4. Source: 2001 Catalan Census.
For 9(a): Sample: Individuals born before 1950 with 50% Most Catalan Surnames of complete population. Number of observations: 1,028,662. Number of surnames: 20,862.   For 9(b): Sample: Individuals born after 1950 with 50% Most Catalan Surnames of of complete population. Number of observations: 1,117,925. Number of surnames: 21,557.

Table 10: ICS over cohorts. 50% Least Frequent Surnames.

(a) Born before 1950 ("Old").

LHS: years of education	(1)	(2)	(3)	(4)	(5)	(6)
CatalanDegreeSurname2		$0.719_{(0.016)}$	$0.453_{(0.016)}$	$0.724_{(0.016)}$		
Surname Dummies			Yes		Yes	
Fake Surnames Dummies				Yes		Yes
Adjusted R-squared	0.2254	0.2270	0.2618	0.2266	0.2612	0.2250
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.972	0.000	0.978

(b) Born after 1950 ("Young").

	(1)	(2)	(3)	(4)	(5)	(6)
CatalanDegreeSurname2		$1.926_{(0.013)}$	$1.073_{(0.014)}$	$1.925_{(0.014)}$		
Surname Dummies			Yes		Yes	
Fake Surnames Dummies				Yes		Yes
Adjusted R-squared	0.0957	0.1120	0.1626	0.1121	0.1583	0.0957
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.414	0.000	0.379

Notes: Regressions as in table 4. Source: 2001 Catalan Census.
For 10(a): Sample: Individuals born before 1950 with 50% with Less Frequent Surnames of complete population. Number of observations: 1,028,665. Number of surnames: 35,464.   For 10(b): Sample: Individuals born after 1950 with 50% with Less Frequent Surnames of complete population. Number of observations: 1,118,983. Number of surnames: 35,284.

Table 11: Assortative Mating in Education & *CatalanDegree* over cohorts

(a) AM in Education

	EduSurname2	
	"Old"	"Young"
EduSurname1	$0.170_{(0.001)}$	$0.303_{(0.001)}$
Observations	2,041,044	2,222,917
R^2	0.3410	0.1997

(b) AM in *CatalanDegree*

	CatDegreeSurname2	
	"Old"	"Young"
CatDegreeSurname1	$0.217_{(0.001)}$	$0.328_{(0.001)}$
Observations	2,041,044	2,222,917
R^2	0.5110	0.2778

Notes: All regressions include gender, age and place of birth dummies. Standard errors in parenthesis. Source: 2001 Catalan Census. Samples: Spanish citizens living in Catalonia with frequency of first and second surname larger than one, born before 1950 ("Old") and after 1950 ("Young"), respectively.

Table 12: ICS with Complete Surnames ("Siblings"). Accumulated frequencies.

(a) Spanish citizens living in Catalonia

LHS: years of education	(1)	(2)	(3)	(4)	(5)	(6)
Adjusted R^2, Surname Dummies	0.5486	0.5416	0.5375	0.5326	0.5283	0.4696
Adjusted R^2, Fake Surnames Dummies	0.3244	0.3224	0.3218	0.3228	0.3232	0.3354
Informational Content of Surnames	0.2242	0.2192	0.2157	0.2098	0.2051	0.1342
Observations	774,788	1,315,853	1,664,717	1,900,652	2,067,590	3,695,479
Number of surnames	387,394	567,749	654,965	702,152	729,975	811,502
Max number of people per surname	2	3	4	5	6	All sample

(b) 50% Most Catalan Surnames

	(1)	(2)	(3)	(4)	(5)	(6)
Adjusted R^2, Surname Dummies	0.5394	0.5355	0.5328	0.5290	0.5254	0.4928
Adjusted R^2, Fake Surnames Dummies	0.3116	0.3104	0.3105	0.3112	0.3122	0.3182
Informational Content of Surnames	0.2278	0.2251	0.2223	0.2178	0.2132	0.1746
Observations	551,164	921,026	1,145,428	1,288,861	1,386,139	1,847,738
Number of surnames	275,583	398,872	454,975	483,662	449,875	536,572
Max number of people per surname	2	3	4	5	6	All sample

Notes: All regressions include gender, age and place of birth dummies. Fake-surnames have the same distribution as Surnames and are allocated randomly. Each column includes a maximum of x people per surname, where $x = \{2, 3, 4, 5, 6, all\}$ for columns (1) to (6), respectively. Source: 2001 Catalan Census.
For 12(a): Sample: Spanish citizens living in Catalonia aged 25 and above, with frequency of complete surname larger than one.
For 12(b): Sample: Individuals with 50% Most Catalan Surnames of sample in table 12(a).

Table 13: ICS with Complete Surnames ("Siblings") over cohorts.

(a) Born before 1950 ("Old")

LHS: years of education	(1)	(2)
Surname Dummies	Yes	
Fake Surnames Dummies		Yes
Adjusted R-squared	0.3664	0.2247
Surnames jointly significant*	Yes	No
(p-value)	0.000	0.248

(b) Born after 1950 ("Young")

LHS: years of education	(1)	(2)
Surname Dummies	Yes	
Fake Surnames Dummies		Yes
Adjusted R-squared	0.3158	0.0995
Surnames jointly significant*	Yes	No
(p-value)	0.000	0.547

Notes: Regressions as in table 12. Source: 2001 Catalan Census. For 13(a): Sample: Individuals born before 1950 of sample in table 12(a), column 6. Number of Observations: 1,606,685. Number of complete surnames: 397,869. For 13(b): Sample: Individuals born after 1950 of sample in table 12(a), column 6. Number of Observations: 1,902,912. Number of complete surnames: 468,995.