Online Appendix

The Informational Content of Surnames, the Evolution of Intergenerational Mobility and Assortative Mating*†

Maia Güell
University of Edinburgh,
FEDEA, CEPR, & IZA

José V. Rodríguez Mora‡
University of Edinburgh & CEPR

Christopher I. Telmer
Tepper School of Business
Carnegie Mellon University

October 2014

*This paper was previously circulated as Güell, Rodríguez Mora, and Telmer (2007): "Intergenerational Mobility and the Informative Content of Surnames," CEPR Discussion Paper 6316.

†We are grateful to Joan Gieseke for editorial assistance.

‡Corresponding author: School of Economics, The University of Edinburgh, 31 Buccleuch Place, Edinburgh EH8 9JT, United Kingdom. Email: sevimora@gmail.com

Abstract

This document is an appendix that accompanies our paper "Intergenerational Mobility and the Informational Content of Surnames." Section 1 provides robustness results that correspond to our paper's baseline model. These results consist of increasing and decreasing the model's fertility variance, income variance and mutation rate. In each case we find that our model's main qualitative results are unchanged. In Section 2 we relax a key assumption of our paper's baseline model, that the surname and income distributions are independent of one another. The key result of this extended model is that surname frequency is informative in and of itself. In Section 3 we provide a set of supplemental empirical results in which we find evidence that surname frequency *is* informative for educational attainment. Subsection 3.1 shows that rare-surname selection bias is not driving our finding of decreased mobility over time. Section 4 provides further details of our calibration exercise. In Section 5 we report further empirical results of our sibling analysis. Finally, Section 6 reports cohort-based results by splitting the population into old (born before 1950) and young (born after 1950).

Contents

1 Robustness of Baseline Model

Our baseline model is that which appears in Section 3 of our paper. Here, we demonstrate that, for our baseline model, the relationship between the inheritance parameter, ρ, and the focal point of our study, the ICS, is robust to different values of the conditional variance of income, the mutation rate of surnames and family size.

In Online-Appendix Figure 1 we plot the equivalent to the R^2 figures from the paper, but with a fertility process with higher family variance. We find no qualitative differences.

Online-Appendix Figure 2 plots analogous figure for an income process where the conditional variance of income is increased by a factor of 10, while Online-Appendix Figure 3 does so for a much smaller value of the conditional variance. The qualitative aspects of the figures are identical to those from our paper. Finally, in Online-Appendix Figures 4 and 5 we show the effects of increasing (decreasing) the mutation rate by a factor of 10. Again, there are no qualitative differences. With larger mutation rates the magnitude of the effects is larger (in particular for low values of ρ), as there are more uncommon surnames, but the qualitative results are the same. Notice that the results are robust *even with very small values of* μ, as this generates enough surname variation.

2 Extended Model: Surname Frequency

There is the possibility of dependence between fertility and income, something we have ruled out in the paper. This section involves relaxing a key assumption of the baseline model of the paper (Section 3), 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 the rest of this section we use simulations 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 magni-*

tude of the relationship depend 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 below. The main result 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.

In section 3 we also report the associated empirical evidence. We find that frequency and educational attainment are indeed related, albeit weakly.

2.1 Modeling Surname Frequency

We now allow our model to have 3 income groups *rich, poor and middle class*, the first two representing the 20% richer and poorer respectively. We assume that the probability of having children and the number of children born differ across these groups. Let $\{q_r, q_m, q_p\}$ be the probability that rich, middle class and poor people give birth, and $\{m_r, m_m, m_p\}$ be the number of children, conditional on giving birth. In order to rule out population growth we impose $\frac{1}{5} \times q_r \times m_r + \frac{3}{5} \times q_m \times m_m + \frac{1}{5} \times q_p \times m_p = 1$. Otherwise, however, the expected number of children, $q_j\, m_j$ can differ across groups. We also allow for differences in surname mutation rates: $\{\mu_r, \mu_m, \mu_p\}$.

An association between the surname and income distributions can now arise for one or more of three reasons: differences in birth probabilities, q_k, average fertility rates, $q_j m_j$, and mutation rates, μ_j. We now examine each in turn.

2.2 Differences in Birth Rates

We refer to differences in q_j — the likelihood of having *sons* — as the *"hereu effect."*[1] They bear directly on the survival rates on surnames, but have no effect on the probability that the size of the surname grows or decreases. Imagine, for instance, a society in which the rich and the poor have the same expected number of children, but the rich have them with certainty while the poor have them stochastically ($q_r = 1$, $m_r = 1$; $\quad q_p = \frac{1}{2}$, $m_p = 2$). Then the probability of lineage survival is 1 for the rich but only 1/2 for the poor. Now suppose that there are 100 surname mutations among the rich and 100 among the poor. After one period the mutations of the rich will all remain, whereas only 50 will remain for the poor (each with two people). Note the key mechanism. The surname death rate is different for different income groups, while the inflow is the same in all of them. The groups with a larger survival

[1] In traditional Catalan society the property of the family farm was inherited by the oldest son (not the daughter), who was called "hereu" (inheritor). The other children would typically be compensated by other forms of education (such as becoming a priest), or by dowry, or with cash. This institution had important consequences relating to the average size of farms (and not letting them become too small), but it had the drawback that families needed a son if they wanted their farm to remain in their lineage. Old-time Catalan farmers seemingly wanted their farms to remain in their lineages, so they wanted sons; having only daughters would not suffice. The way to ensure this was to keep having children at least until a boy was born. The probability of a family's lineage dying was very low if they had a farm, because at least the male child would continue to keep the lineage alive. Families without farms would be less concerned with having a male child, and thus the probability of disappearance of the lineage would be higher.

rate are bound to accumulate a larger number of infrequent surnames.

Online-Appendix Figures 6, 7 and 8 report the results of a simulation in which everyone has the same expected number of children ($q_j m_j = 1 \quad \forall j$) but where the rich *always* have a male child, so that $q_r = 1; \quad m_r = 1$, while for the middle class $q_m = 1/2$ and $m_m = 2$ and for the poor $q_r = 1/4$ and $m_r = 4$. There are three main points.

1. Online-Appendix Figure 6 shows that the frequency of the surname is informative: a higher frequency is associated with less income. Also the more important is inheritance (*i.e.*, the larger is ρ), the larger is the absolute value of the t-statistic of the frequency. This second feature is particularly important.

   To understand this, imagine two mutations. The first occurs among the rich, giving birth to the lineage *Richmanson*. The second occurs among the poor, giving birth to the lineage *Poormanson*. Now, suppose that the degree of inheritance is large. The lineage *Richmanson* will survive for a long time *and will have a small frequency during that time*. This is a consequence of high income persistence, implying that the sons of *Richmanson* will remain rich, have sons of their own, and thus continue the lineage. Also, although the surname will not disappear, it is also unlikely that the surname's frequency will grow. This is because the rich do have sons, although not many.

   On the other hand, it is unlikely that the lineage *Poormanson* will survive *and remain infrequent*. *Poormanson* and *Richmanson* have the same expected number of sons, but *Poormanson* has a higher variance.

He is more likely to have no sons (thus ending his lineage), but if he does have sons he will have more than the *average* rich man. As a consequence infrequent surnames will tend to belong to rich people, and only seldom will a poor man have an uncommon surname.

If the degree of inheritance were smaller one would not see such a large frequency effect, as lineages that began rich have a larger probability of becoming poor (and then disappearing). There would be less concentration of rich people with infrequent surnames.

2. Online-Appendix Figure 7 shows that the distribution of surnames *does* depend on the income process. This stands in sharp contrast to our previous results with no link between demographics and income. The distribution of surnames, being well approximated by a geometric distribution, is characterized by the number of people per surname and the Gini index of the surname distribution. The number of agents per surname decreases with the degree of inheritance, while the Gini index increases. The reason for the first is that if inheritance is very important (high ρ) rich individuals tend to have one-of-a-kind surnames. Once they get the surname it only changes if there are mutations, but its frequency does not grow. The Gini index is large because a few surnames (those of the poor) hold a large percentage of the population. The distribution becomes very skewed.

3. Finally, Online-Appendix Figure 8 shows that our logic from the simpler model carries forth here. When conditioning on the specific surname, and thus approximating family relationships, the ICS increases

with ρ in the same manner as it did before. The mechanism of grouping siblings together (surnames being an informative partition of the population, as it relates to family) is still working. This will be important for our empirical approach: irrespective of the informativeness of frequency, we can infer the degree of mobility by looking at the ICS alone.

2.3 Differences in Average Fertility

Differences in average fertility between income groups (differences in $m_j q_j$) are more complex.[2] This is because they affect both the survival probability of a lineage and the rate of change of its frequency, conditional on surviving. Differences in average fertility also change the relative population holding the surnames. That is, suppose that the rich have larger average fertility. Then not only do they have a lower probability of lineage extinction (and a high incidence of infrequent surnames), but this will also induce the rich surnames to become frequent relatively quickly. The key to determine if a infrequent surname is going to indicate wealth or its absence is the interaction between m_j and q_j.

Notice finally that by inducing differences in reproductive patterns between rich and poor individuals, the unconditional distribution of income in the population will *not* be the same as the unconditional distribution of

[2]Note that we refer to *males* here, the average number of (reproductively capable) *male* offspring that a *male* adult has. The correlation between "male fertility" and income can go in exactly the opposite direction from female fertility. Educated females are known to have fewer children than uneducated ones, but that is not necessarily the case for males. It is not uncommon for successful males to have children with more than one female, by either remarriage, polygamy or out-of-wedlock relationships.

income from our baseline model. For instance, if the average fertility of the rich is relatively large, then a positive income shock in one generation will transmit to more individuals (on average) than a negative one of the same magnitude. The income distribution would switch toward higher levels of average income.

Below we present the result of simulations with differences in average fertility. We show that the ICS maintains its monotonous relationship with inheritance, as surnames are still approximating recent common ancestry. The relationship between frequency and inheritance is very complex (sometimes positive, sometimes negative). The relationship between ICS and inheritance is stable, clear, always increasing and positive. This lends credence to our emphasis on the ICS in our empirical work.

In Online-Appendix Figures 9, 10 and 11 we show the results of a simulation that the only result that changes with respect to our benchmark simulation is that the expected number of children differs among the income groups (even if the probability of having male offspring is the same for all of them, $q_j = \frac{1}{2} \quad \forall j$). Let E_j be the expected number of children for income group j, where $E_j = q_j \times m_j$. In this simulation $E_r = 1.5; \quad E_m = 1; \quad E_p = \frac{1}{2}$.

In Online-Appendix Figure 9, we observe that the t-statistic of frequency is always positive, significantly different from zero, and increases with inheritance, the reason being that rich people have more children, which makes surnames more common.

Notice also in this case that the distribution of surnames is affected by inheritance. In Online-Appendix Figure 10 more inheritance implies a larger Gini index and a smaller number of surnames per person. This is because

with more inheritance rich people lineages become large. Of course they can not be all rich (as the definition of "rich" and "poor" is relative), so the less fortunate between them moves down to lower incomes. *Their* lineages do not disappear, even if the probabilities of having male descendants decrease substantially, as their rich cousins share their surname with them. The mutations that happen among the poor would be short living, the mutations among the rich will survive by making their surname large.

Finally in Online-Appendix Figure 11 we meet again with our main result. Irrespectively of if frequency of the surname is positively (as in this case) or negatively (as in the previously) associated to inheritance, it is always the case that more inheritance translates into a larger informative content of surnames. This is because ICS refers to family bonds, while frequency has information because the shape of the distribution of surnames is a function of income distribution once lineage birth/death probabilities depend on the income of the agents.

2.4 Differences in the Mutation Rate

It is straightforward to see that frequency of the surname has information on the income of its holder if there are differences in the rates of birth of lineages associated with income differentials. The reasons are basically the same as those given above. Suppose, for example, that rich mutate their names more frequently than the poor. Then the inflow of new lineages would be larger among the rich than the poor and the infrequent surnames would tend to belong to the rich. The opposite would happen if the poor mutated their

names more often.

The predicted relationship between frequency and income, then, depends on which way the mutation-rate differentials go. Empirically, there are countervailing effects. On one hand there are reasons to believe that surname mutations are more likely to occur among the rich. The number of hyphenations, and even the sheer length of the surname are probably associated to higher income, as rich people may like to signal their status through their surnames. This could well work in a form akin to first (given) name allocation. It is well known that the better-off choose names for their offspring that are new, and different from the most common ones in their society (*c.f.*, Fryer and Levitt (2004) and Levitt and Dubner (2005)).

On the other hand, migration is probably the most common form of introducing new surnames into a given population, and in our context it could be interpreted as mutations. Emigrants tend to be poor. They also tend to have surnames that from the point of view of the recipient population are unusual. Most often they are simply unique because the possibility of mutation is very likely to increase a lot as a direct consequence of migration. Transliteration of foreign scripts and alphabets, orthographic and phonetic differences between countries all this adds up to generate new surnames that are new not only from the point of view of the recipient populations, but also in the original population of the migrant.

An additional complication is that the relationship between migration and mutation depends on the difference between the surname distribution of the origin and recipient populations. A migrant from Morocco to Spain is more likely to introduce a new surname in Spain than a migrant from

Ecuador. In the same manner, if migration happens between regions that are "close" from a surname distribution point of view the number of observed mutations will be lower than if the regions are far apart.

To conclude this subsection, we find that (i) there are reasons to expect that the surname distribution should be a function of the income process, (ii) characteristics of the surname such as its frequency are — in addition to the specific surname itself — likely to be informative for economic well-being, (iii) there are many forces at work, often going in different directions, and (iv) the ICS measure seems robust to these issues for the study of the importance of inheritance.

3 Empirical Results on Surname Frequency

In Online-Appendix Table 1 we regress the the *frequency* of an individual's first surname. on his education.[3] The role of the second surname, as before, is to control for ethnicity using the $CatalanDegree$ variable. The negative point estimate on the frequency variable implies that a lower frequency is associated with a higher level of educational attainment (after controlling for ethnicity), see columns 1 and 2. Columns 3 and 4 show that the frequency of fake-surnames is not significant. Specifically, the value of -23.696 implies that a one standard deviation increase in frequency translates into 0.15 fewer years education. This is a decrease of 3% of one standard deviation of the

[3]It is important to understand that this is fundamentally different what we did in Section 6.2 of our paper. There, infrequent surnames were shown to be informative simply because they are associated with familial linkages. This was just as true for the highly educated as for the poorly educated. Here, we ask whether or not the *frequency itself* is correlated with educational attainment.

level of education.[4]

This result indicates that either the death rate of lineages is smaller among the more educated, or their birth rate is larger, or both. Either effect is quite conceivable. The newly rich, for instance, are more likely to create new surnames (by hyphenation of first and second usually). It can also be related to an "hereu" effect inducing better-off families to have children until the point of insuring one male descendant. We discuss this further below. Similarly, this is what we would expect to see if educated *males* have more children that non-educated *males*, perhaps because they are more likely to form additional families after divorce. Note that, we are excluding foreign immigrants and if we were to include them the results could very well change, as the effective mutation rate for the poorly educated would be much larger.

3.1 Time Evolution of Rare-Surname Selection Bias

As discussed above, Online-Appendix Table 1 establishes one important sense in which rare surname holders are not a random selection of the population. But, as section 7 of our paper shows, this fact does not affect our estimate of the inheritance parameter ρ. The next question is, if this selection is the primary reason for an increasing ICS (as opposed to declining mobility). In this section we argue that the answer to this question is no.

Online-Appendix Figure 12 plots the mean and standard deviation of educational attainment for the complete population as well as the segment of the population with the 50% least-frequent surnames.

[4]For the sample of the table the mean of frequency of surname 1 is 0.00327 and its standard deviation 0.00620.

The figures confirm the regression evidence from Online-Appendix Table 1; people with infrequent surnames are indeed more highly educated. But this is true for all of our cohorts and in a similar magnitude. This suggests that our ICS results are not driven by the differences in the composition of rare surnames over time.

4 Further Details of the Calibration

In this section we numerically demonstrate the properties of the joint distribution of surnames and income that we use in section 7 of our paper and provide further details of the calibration process.

Remember that the extended model consists on 5 parameters:

(i) ρ, the correlation of income of parents and children.

(ii) $\sqrt{V_\epsilon}$, the conditional standard deviation of educational attainment.

(iii) μ, the mutation rate of surnames.

(iv) m the number of sons that a father has if it has children.

(v) d, the differences in fertility depending on education.

Notice that the baseline model of section 3 of our paper just adds the additional restriction that there are no differences in fertility between education groups. Thus, it is included in our analysis.

We set V_ϵ so that the unconditional standard deviation matches that of our data. To determine suitable values for the rest of the parameters we generate a (very large) grid of parameter values, and generate an artificial economy for each element of the grid. Notice that the baseline model is included, as it simply consists on assuming that there are no differences in

fertility for different educational groups.

Each of the artificial economies (each combination of the 4 parameters plus the adequate $\sqrt{V_\epsilon}$) an "artificial census" is generated in an economy with 1.5 million individuals. Each census consists of a set of surname and education level for each individual. The program runs 125 iterations (to get rid of initial conditions), then an artificial census is collected during 25 further iterations.[5]

For each "artificial census" of the grid we calculate the value of four moments for which we have empirical counterparts in the data: (i) The ICS, calculated in each artificial economy in the same manner than in Section 6 of our paper is calculated with real data. (ii) The GINI index and the (iii) average number of persons per surname (PPS) of the surname distribution, calculated in the same manner than with the actual census. And finally, (iv) the coefficient of surname frequency in a regression explaining educational attainment, as performed with the actual census in Section 3.

Online-Appendix Figure 13 summarizes the results of this exercise, and demonstrates the following properties:

1. **The Gini coefficient is hump-shaped in the mutation rate, μ. The value of μ that maximizes the Gini coefficient is (essentially) independent of both ρ and m.**

   Online-Appendix Figures 13(a), 13(b) and 13(c) show the value of the GINI index as a function of μ and d, m and ρ respectively, with the value of the other two variables (ρ and m for Online-Appendix Figure

[5]Software to generate the grid, as well as all the other software needed to interpret it, is available from the authors upon request.

13(a), and so on) are fixed at their calibrated value (from the second column of table 5 in our paper). Clearly, the maximum is achieve by a value of μ which is independent of the value taken by the other variables. Further numerical checks (more on this bellow) insure us that this value is fix for all possible combinations of the other parameters. As we report in the main text, this maximum falls short of the observed Gini for all parameter values, so we choose the value of μ as the one that sets the Gini index as good as possible. This is, μ of 0.0067.

2. **The value of the ICS is independent of both m and d. It increases with either ρ and μ**

   Figure 13(d) shows that value of the ICS for all combinations of m and d while we keep fixed the calibrated values of ρ and μ. It is clearly flat. Indicating that neither m or d are of any consequence for the determination of the ICS.

   Figure 13(e) shows the value of the ICS for all combinations of ρ and μ while we keep fixed the calibrated values of m and d. Clearly the ICS increases with both ρ and μ.

   Imagine that you were going to keep constant the ICS at a certain level, cutting this surface horizontally across a certain value of the ICS. The set of values of rho and μ that keep the ICS fixed a that level establishes a decreasing relationship between these variables: a fixed ICS can be achieved by either a large ρ and low μ, or a low ρ and a large μ. Bellow we draw this relationship for the value of the ICS that we obtain from the data. Notice that this implies that given the value

of μ that maximizes the GINI of the surname distribution, there is only one value of ρ that sets the ICS right. This is the calibrated value of ρ.

A second interesting exercise is to imagine that we cut through the surface in Figure 13(e) across a fixed value of μ. In such a case we obtain an increasing relationship between ICS and ρ. This is the same relationship that we saw in Figure 2 of our paper in the main model (Section 3).

3. **Given a certain value of μ, the average number of persons per surname (PPS) is independent of ρ and d. Thus, given a value of μ there is only a value of m that matches the observed PPS**

   Figure 13(f) shows how any combination of μ and m map into PPS (for the fixed calibrated values of ρ and d). We notice that given any value of PPS, and a value of μ there is only one compatible value of m, which is then determined. Figures 13(g) and 13(h) show that this value is not sensitive at all to either ρ or d

Thus, it is possible to calibrate recursively the values of μ, m and ρ. The value of d is obtained then as a residual in order to match the coefficient of frequency obtained from the data.

A way of showing the robustness of the parametrization ρ (our variable of interest), it is to look at the remaining two figures.

Figure 13(i) plots the sets of pairs m and μ that maximize the value of the GINI *for all possible combinations of the other parameters.* This is, each line in the graph represents the values of m and μ that generate the crest of the surface 13(b); and we draw this line for all possible combinations of ρ

and d (be then the calibrated parameters or not). What we observe is that the set of reasonable values for μ is very narrow.

Figure 13(j) plots the sets of pairs of ρ and μ that match the ICS from the data *for all possible combinations of the other parameters*. This is, each line in the graph represents the values of ρ and μ that would be across the line generated by cutting horizontally the surface 13(e) at the height of the ICS observed in the data; and we draw this line for all possible combinations of m and d (be then the calibrated parameters or not). Clearly this relationship is essentially independent of m and d.

Given that the set of reasonable values for μ is very narrow, the set of empirically compatible values for ρ is itself very narrow, and independent of the model configuration. The value for ρ is systematically of around 0.6.

5 Analysis of Sibling: Further Results

In this section we provide some further results from our sibling analysis.

In our paper (section 9), to approximate siblings as much as possible we concentrate on those who share their complete–surname by one or two other persons. Online-Appendix Figure 14 is as Figures 9 and 10 in our paper but for the *whole population*. As can be seen, our results do not change, the same qualitative pattern arises over time.

We next present the values of our sibling correlation proxy (SCP). Online-Appendix Table 2(a) reports the SCP for those who share their complete–surname by one or two other persons as well as for the whole population. We notice two facts. First, the SCP declines as we increase the likelihood

of spuriously grouping together individuals who are not siblings. Second, these correlations are much higher than the ICS reported in our paper. The reason is that the SCP is a very different (and much finer) partition of the data that approximates very closely the sibling relationship. While the ICS is a more coarse partition that is informative on family relationship broadly understood. Since the SCP and the ICS are based on very different partitions of the population, their values are not directly comparable. But the time trend of these two measures of mobility are comparable, as discussed in Section 9 of our paper.

Online-Appendix Table 2(a) cannot control for ethnicity because our CatalanDegree variable and the complete–surname dummy are based on a common surname. Online-Appendix Table 2(b) therefore reports on a sub-population containing only the 50% most Catalan surnames, resulting in a more ethnically homogeneous population. Our results do not change.

6 Cohort-Based Empirical Results: Old and Young

In this section we report cohort-based results by splitting the population into old (born before 1950) and young (born after 1950).

6.1 Cohort-Based ICS

In this section we report our dynamic, cohort-based results (Section 8 of our paper) by splitting the population into old (born before 1950) and young

(born after 1950).

Online-Appendix Table 3(a) reports results for the same regressions as Table 2 in the paper, except that the population is restricted to those born *before* 1950.[6] The results are similar to those for the entire population. Online-Appendix Table 3(b) includes only those born *after* 1950. There are three notable results. First, the explanatory power of the regressors is much lower. Not surprisingly, 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. Finally — and most importantly — the ICS is substantially higher for younger cohort than for the older cohort.

We now present a battery of robustness checks, basically replicating the checks that we undertook for the single cross-section in Section 6.1 of our paper.

Online-Appendix Table 4 replicates the results of Online-Appendix Table 3 but restricting the sample to the 50% of the population with the most Catalan surnames. As before, the ICS increases, 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 4 and Figure 4 from Section 6.2 of our paper provided a powerful confirmation of our model by showing that the ICS is much larger when we

[6]The data include both people born in and outside of Catalonia. If we were to exclude the latter (as in Table 3b in the paper), we would include the children of the immigrants in the population of 'young' but not their parents in the population of the 'old'. Thus, to look at time trends could be misleading.

consider only (relatively) rare surnames. Online-Appendix Table 5 replicates the analysis for age-cohorts, excluding individuals with surnames that are in the upper 50% of the commonality distribution. Again, the ICS increases when we exclude names that, almost by definition, cannot be informative about familial linkages.

6.2 Cohort-Based Sibling Correlations

In this section we report our siblings cohort-based results (Section 9 of our paper) by splitting the population into old (born before 1950) and young (born after 1950).

Online-Appendix Table 6 shows the tables for young and old using complete–surnames. As before, our Sibling Correlation Proxy (SCP) increases across the old and young cohorts.

6.3 Cohort-Based Assortative Mating

In this section we report our assortative mating cohort-based results (Section 10.2 of our paper) by splitting the population into old (born before 1950) and young (born after 1950).

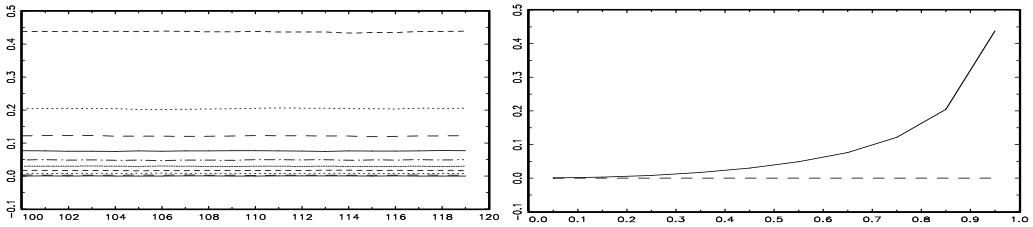
Online-Appendix Table 7(a) reports results for the education characteristic. We see that the correlation between the educational dimension of first and second surnames increases from the old cohort the young cohort. Educational assortative mating seems to have increased. Online-Appendix Table 7(b) reports the analogous measurement for ethnicity. The correlation also increases. The parents of younger cohorts seem more likely to have married

within their ethnic background than the parents of the older cohorts.

References

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

Levitt, S. and S. Dubner (2005). *Freakonomics*. HarperCollins.

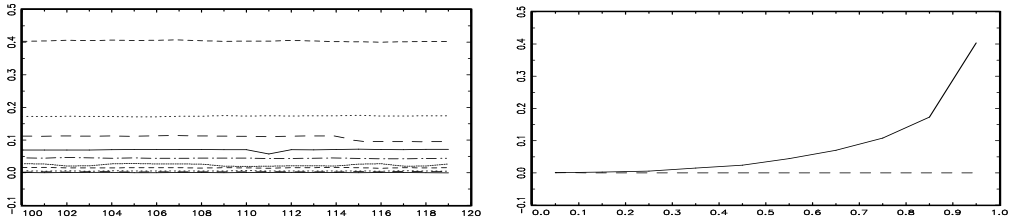


(a) Time series of R_L^2 for different values of ρ

(b) Average R_L^2 (solid line) and R_F^2 (dotted line) against ρ

Online-Appendix Figure 1: High Family Size Variance

Notes: Model Simumations with Parameter Values: N_0=1000000; V_ε=1.000; μ=0.0200; q=0.25; m=4; $\rho \in [0.05, 0.95]$.

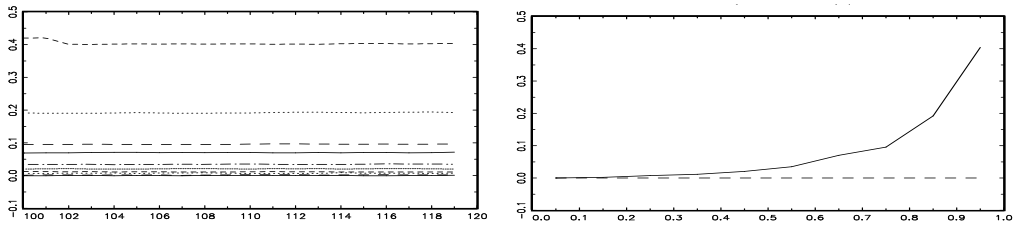


(a) Time series of R_L^2 for different values of ρ

(b) Average R_L^2 (solid line) and R_F^2 (dotted line) against ρ

Online-Appendix Figure 2: Differences in V_ε: High Conditional Variance

Notes: Model Simumations with Parameter Values: N_0=1000000; V_ε=10.000; μ=0.0200; q=0.50; m=2; $\rho \in [0.05, 0.95]$.

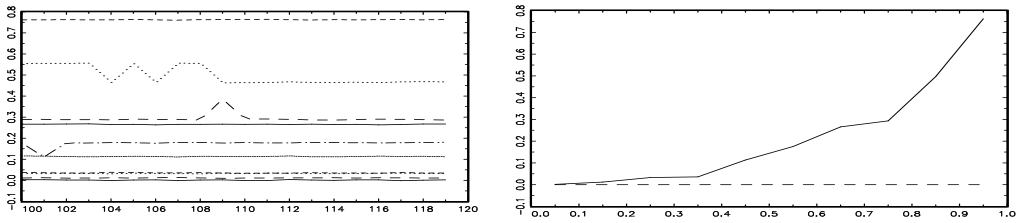


(a) Time series of R_L^2 for different values of ρ

(b) Average R_L^2 (solid line) and R_F^2 (dotted line) against ρ

Online-Appendix Figure 3: Differences in V_ε: Low Conditional Variance

Notes: Model Simumations with Parameter Values: N_0=1000000; V_ε=0.100; μ=0.0200; q=0.50; m=2; $\rho \in [0.05, 0.95]$.

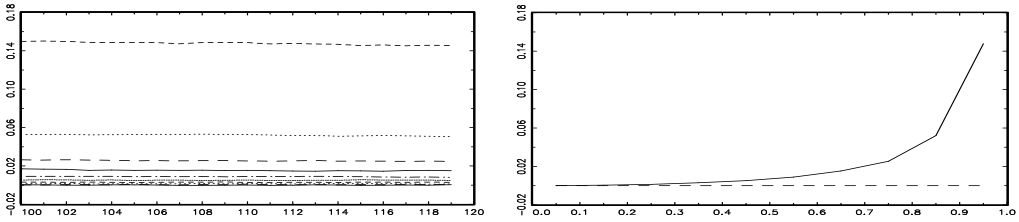


(a) Time series of R^2_L for different values of ρ

(b) Average R^2_L (solid line) and R^2_F (dotted line) against ρ

Online-Appendix Figure 4: Differences in μ: High Mutation Rate

Notes: Model Simumations with Parameter Values: N_0=1000000; V_ε=1.000; μ=0.2000; q=0.50; m=2; $\rho \in [0.05, 0.95]$.



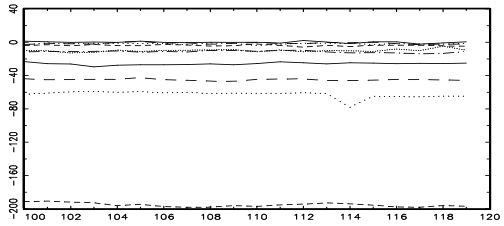
(a) Time series of R^2_L for different values of ρ

(b) Average R^2_L (solid line) and R^2_F (dotted line) against ρ

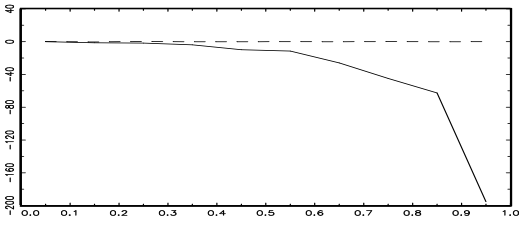
Online-Appendix Figure 5: Differences in μ: Low Mutation Rate

Notes: Model Simumations with Parameter Values: N_0=1000000; V_ε=1.000; μ=0.00200; q=0.50; m=2; $\rho \in [0.05, 0.95]$.

Online-Appendix Figures 6 to 8: "Hereu Effect": Differences across income groups in the probability of survival of surnames

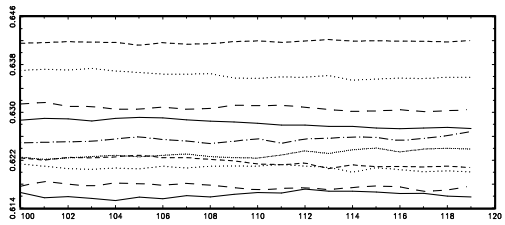


(a) Time series of t-statistic of real surname frequency for different values of ρ

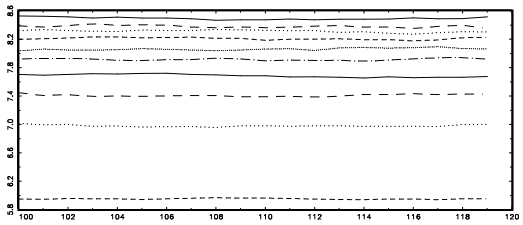


(b) Average t-statistic of real surname frequency (solid line) and fake surname frequency (dotted line) against ρ

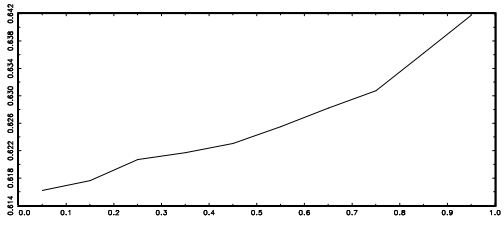
Online-Appendix Figure 6: "Hereu Effect": surname frequency



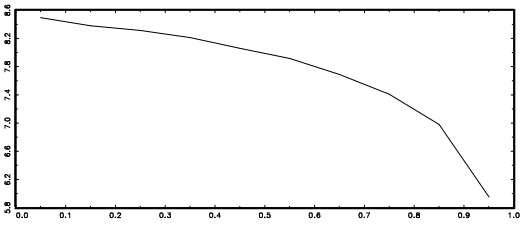
(a) Time series of Gini of surname distribution for different values of ρ



(b) Time series of average number of agents per surname for different values of ρ

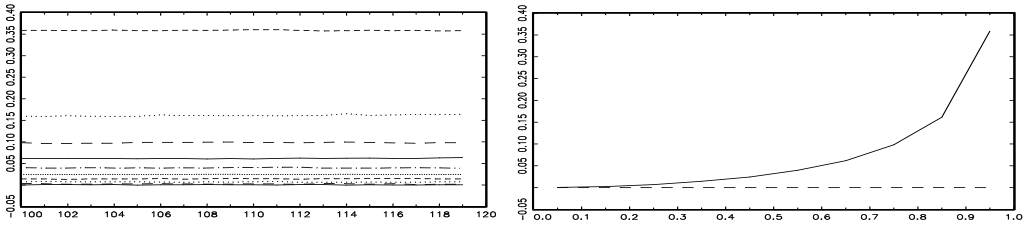


(c) Average Gini of surname distribution against ρ



(d) Average average number of agents per surname against ρ

Online-Appendix Figure 7: "Hereu Effect": surname distribution



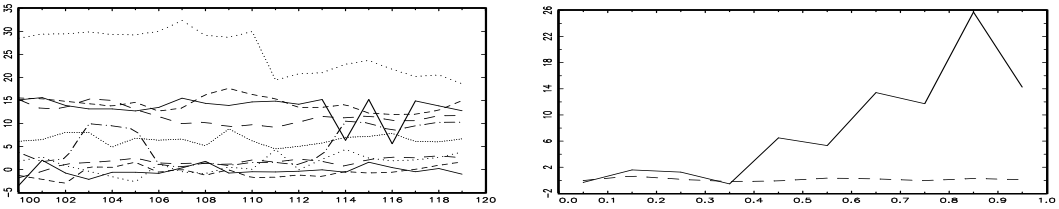
(a) Time series of R^2_L for different values of ρ

(b) Average R^2_L (solid line) and R^2_F (dotted line) against ρ

Online-Appendix Figure 8: "Hereu Effect": adjusted R^2

Notes for Online-Appendix Figures 6 to 8: Model Simumations with Parameter Values: N_0=1000000; V_ε=1.000; μ=0.0200; $q_j = \{1.00, 0.50, 0.25\}$; $m_j = \{1.00, 2.00, 4.00\}$ where $j = \{r, m, p\}$; $\rho \in [0.05, 0.95]$.

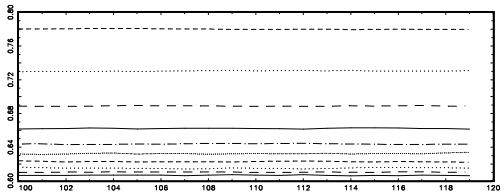
Online-Appendix Figures 9 to 11: Fertility differences: Differences across income groups in the average fertility



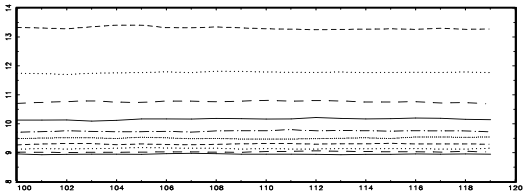
(a) Time series of t-statistic of real surname frequency for different values of ρ

(b) Average t-statistic of real surname frequency (solid line) and fake surname frequency (dotted line) against ρ

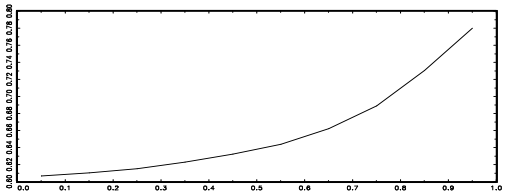
Online-Appendix Figure 9: Fertility differences: surname frequency



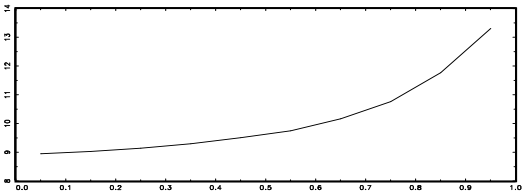
(a) Time series of Gini of surname distribution for different values of ρ



(b) Time series of average number of agents per surname for different values of ρ

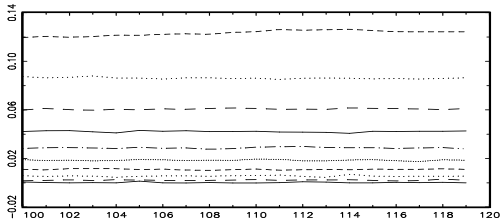


(c) Average Gini of surname distribution against ρ

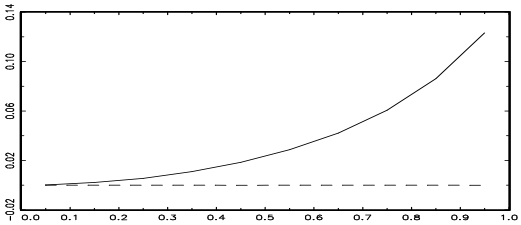


(d) Average average number of agents per surname against ρ

Online-Appendix Figure 10: Fertility differences: surname distribution



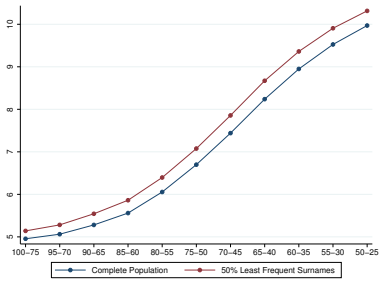
(a) Time series of R^2_L for different values of ρ



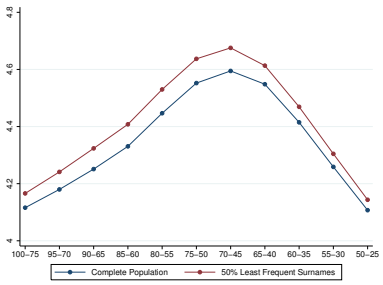
(b) Average R^2_L (solid line) and R^2_F (dotted line) against ρ

Online-Appendix Figure 11: Fertility differences: adjusted R^2

Notes for Online-Appendix Figures 9 to 11: Model Simumations with Parameter Values: N_0=1000000; V_ε=1.000; μ=0.0200; $q_j = \{0.50, 0.50, 0.50\}$; $m_j = \{3.00, 2.00, 1.00\}$ where $j = \{r, m, p\}$; $\rho \in [0.05, 0.95]$.



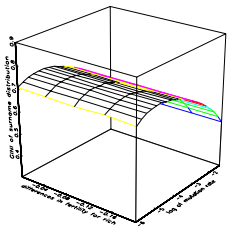
(a) Average



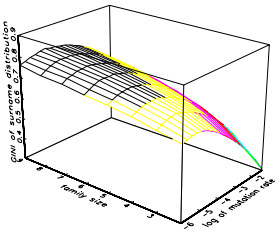
(b) Standard Deviation

Online-Appendix Figure 12: Evolution of years of education over moving windows of cohorts

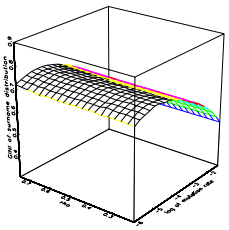
Notes: Overlapping age-cohorts are described in caption to Figure 5 of our paper. Source: 2001 Catalan Census.



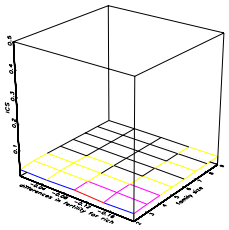
(a) GINI, μ and d



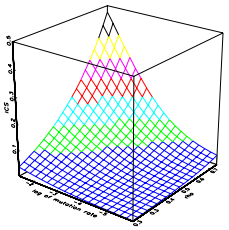
(b) GINI, μ and m



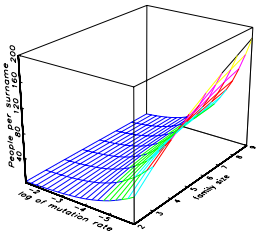
(c) GINI, μ and ρ



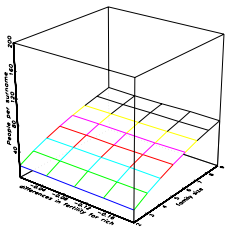
(d) ICS, m and d



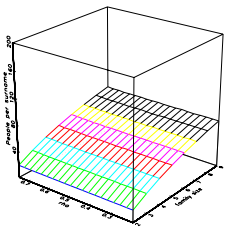
(e) ICS, ρ and μ



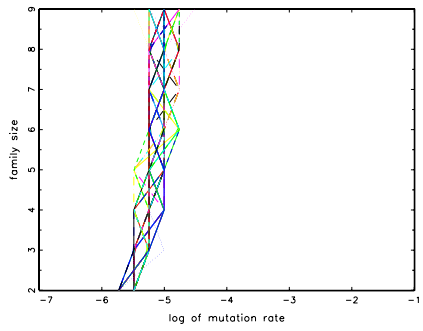
(f) PPS, m and μ



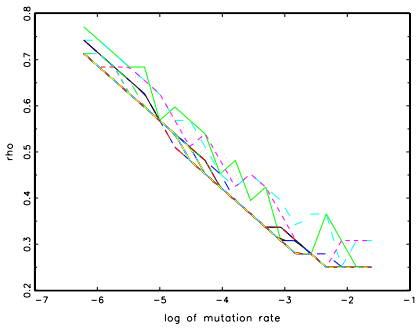
(g) PPS, m and d



(h) PPS, m and ρ



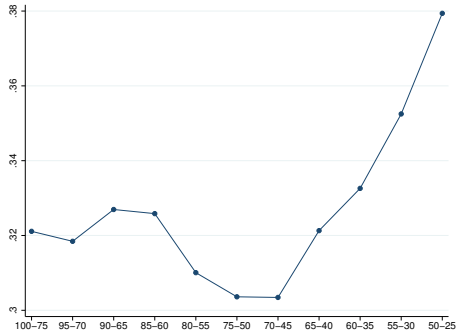
(i) Pairs of m and μ that get closest to GINI for all combinations of ρ and d



(j) Pairs of ρ and μ that match ICS for all combinations of m and d

Online-Appendix Figure 13: Calibration

Notes: Each of the Figures 13(a) to 13(h) plots the values of a moment of the distribution of the artificial economy for all possible combinations of two parameters while keeping the other two parameters constant at their calibrated value. Figure 13(i) plots the combinations of μ and m that get a highest GINI index of the distribution of surnames for all possible combinations of the other parameters. This is, each line is the μ that gets a highest GINI for each m *given a specific value* of ρ and d. Figure 13(j) plots the combinations of μ and ρ that match the data ICS for all possible combinations of the other parameters. This is, each line is the ρ that matches the ICS of the data for each μ *given specific values* of m and d.



Online-Appendix Figure 14: Evolution of Sibling Correlation Proxy, SCP over moving windows of cohorts. All Population.

Notes: Regressions as in Online-Appendix Table 2(a), column (3). Overlapping age-cohorts are described in caption to Figure 5 of our paper. Source: 2001 Catalan Census.

Online-Appendix Table 1: Education and Surname Frequency

LHS: years of education	(1)	(2)	(3)	(4)
FrequencySurname1	-31.256$_{(0.460)}$	-24.625$_{(0.460)}$		
FrequencyFakeSurname1			-0.374$_{(0.448)}$	-0.395$_{(0.445)}$
CatalanDegreeSurname2		1.647$_{(0.011)}$		1.706$_{(0.011)}$
Adjusted R^2	0.2669	0.2745	0.2653	0.2735

Notes: All regressions include age and place of birth dummies. Fake-surnames have the same distribution as Surnames and are allocated randomly. Standard errors in parenthesis. Population: Baseline population. Number of observations: 2,057,134. Source: 2001 Catalan Census.

Online-Appendix Table 2: Sibling Correlation Proxy, SCP.

(a) Spanish citizens living in Catalonia

LHS: years of education	(1)	(2)	(3)
Adjusted R^2, Complete–Surname Dummies	0.5025	0.4884	0.4035
Adjusted R^2, Complete–Fake–Surnames Dummies	0.2517	0.2557	0.2664
Sibling Correlation Proxy (SCP)	0.2508	0.2327	0.1371
Observations	428,134	655,303	1,487,191
Number of Complete–Surnames	214,067	289,790	374,256
Max number of People per Complete–Surname	2	3	All Population

(b) 50% Most Catalan Surnames

LHS: years of education	(1)	(2)	(3)
Adjusted R^2, Complete–Surname Dummies	0.5029	0.4904	0.4390
Adjusted R^2, Complete–Fake–Surnames Dummies	0.2446	0.2434	0.2525
Sibling Correlation Proxy (SCP)	0.2583	0.2470	0.1865
Observations	302,486	453,219	743,595
Number of Complete–Surnames	151,243	201,489	234,472
Max number of People per Complete–Surname	2	3	All Population

Notes: All regressions include age and place of birth dummies. Fake–complete–surnames have the same distribution as complete–surnames and are allocated randomly. Population: Male Spanish citizens living in Catalonia aged 25 and above, with frequency of complete–surname larger than one. Source: 2001 Catalan Census.

Online-Appendix Table 3: ICS over cohorts. Baseline population.

(a) Born before 1950 ("Old")

LHS: years of education	(1)	(2)	(3)	(4)	(5)	(6)
CatalanDegreeSurname2		0.972(0.019)	0.635(0.02)	0.965(0.019)		
Surname Dummies			Yes		Yes	
Fake Surnames Dummies				Yes		Yes
Adjusted R^2	0.2063	0.2086	0.2328	0.2083	0.2319	0.2060
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.457	0.000	0.394

(b) Born after 1950 ("Young")

LHS: years of education	(1)	(2)	(3)	(4)	(5)	(6)
CatalanDegreeSurname2		2.049(0.014)	1.218(0.015)	2.045(0.014)		
Surname Dummies			Yes		Yes	
Fake Surnames Dummies				Yes		Yes
Adjusted R^2	0.0763	0.0936	0.1276	0.0938	0.1225	0.0766
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.083	0.000	0.052

Notes: Regressions as in table 2 of the paper. For 3(a): Number of observations: 937,441. Number of surnames: 28,944. For 3(b): Number of observations: 1,119,693. Number of surnames: 29,586. Source: 2001 Catalan Census.

Online-Appendix Table 4: ICS over cohorts. 50% Most Catalan Surnames

(a) Born before 1950 ("Old")

LHS: years of education	(1)	(2)	(3)	(4)	(5)	(6)
CatalanDegreeSurname2		0.676(0.023)	0.442(0.023)	0.685(0.023)		
Surname Dummies			Yes		Yes	
Fake Surnames Dummies				Yes		Yes
Adjusted R^2	0.1896	0.1911	0.2205	0.1915	0.2199	0.1899
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.041	0.000	0.052

(b) Born after 1950 ("Young")

LHS: years of education	(1)	(2)	(3)	(4)	(5)	(6)
CatalanDegreeSurname2		1.688(0.018)	1.003(0.019)	1.688(0.018)		
Surname Dummies			Yes		Yes	
Fake Surnames Dummies				Yes		Yes
Adjusted R^2	0.0652	0.0799	0.1206	0.0794	0.1160	0.0647
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.159	0.000	0.17

Notes: Regressions as in table 2 of the paper. For 4(a): Number of observations: 468,721. Number of surnames: 17,422. For 4(b): Number of observations: 559,847. Number of surnames: 18,471. Source: 2001 Catalan Census.

Online-Appendix Table 5: ICS over cohorts. 50% Least Frequent Surnames

(a) Born before 1950 ("Old")

LHS: years of education	(1)	(2)	(3)	(4)	(5)	(6)
CatalanDegreeSurname2		0.774(0.025)	0.46(0.026)	0.781(0.026)		
Surname Dummies			Yes		Yes	
Fake Surnames Dummies				Yes		Yes
Adjusted R^2	0.2024	0.2040	0.2442	0.2045	0.2437	0.2029
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.837	0.000	0.853

(b) Born after 1950 ("Young")

LHS: years of education	(1)	(2)	(3)	(4)	(5)	(6)
CatalanDegreeSurname2		1.826(0.019)	1.020(0.021)	1.826(0.019)		
Surname Dummies			Yes		Yes	
Fake Surnames Dummies				Yes		Yes
Adjusted R^2	0.0740	0.0893	0.1372	0.0890	0.1332	0.0737
Surnames jointly significant*			Yes	No	Yes	No
(p-value)			0.000	0.159	0.000	0.172

Notes: Regressions as in table 2 of the paper. For 5(a): Number of observations: 468,720. Number of surnames: 28,581. For 5(b): Number of observations: 560,265. Number of surnames: 29,276. Source: 2001 Catalan Census.

Online-Appendix Table 6: Sibling Correlation Proxy, SCP over cohorts.

(a) Born before 1950 ("Old")

LHS: years of education	(1)	(2)	(3)
Adjusted R^2, Complete–Surname Dummies	0.4346	0.4207	0.3350
Adjusted R^2, Complete–Fake–Surnames Dummies	0.1932	0.1944	0.1952
Sibling Correlation Proxy (SCP)	0.2414	0.2263	0.1398
Observations	200,938	296,827	586,136
Number of Complete–Surnames	100,469	132,432	164,226
Max number of People per Complete–Surname	2	3	All Population

(b) Born before 1950 ("Young")

LHS: years of education	(1)	(2)	(3)
Adjusted R^2, Complete–Surname Dummies	0.4057	0.3911	0.2762
Adjusted R^2, Complete–Fake–Surnames Dummies	0.0667	0.0709	0.0773
Sibling Correlation Proxy (SCP)	0.3390	0.3202	0.1989
Observations	261,168	388,950	778,329
Number of Complete–Surnames	130,584	173,178	215,004
Max number of People per Complete–Surname	2	3	All Population

Notes: Regressions as in table 2(a). Source: 2001 Catalan Census.

Online-Appendix Table 7: Assortative Mating in Education & *CatalanDegree* over cohorts

(a) AM in Education

	EduSurname2	
	"Old"	"Young"
EduSurname1	$0.160_{(0.001)}$	$0.274_{(0.001)}$
Observations	920,933	1,105,484
R^2	0.297	0.17

(b) AM in *CatalanDegree*

	CatDegreeSurname2	
	"Old"	"Young"
CatDegreeSurname1	$0.219_{(0.001)}$	$0.330_{(0.001)}$
Observations	920,933	1,105,484
R^2	0.5110	0.278

Notes: All regressions include age and place of birth dummies. Standard errors in parenthesis. Population: Cohorts of male Spanish citizens living in Catalonia aged 25 and above, with frequency of first and second surname larger than one. Source: 2001 Catalan Census.