Brain Drain, Numeracy and Skill Premia during the Era of Mass Migration: Testing the Roy-Borjas Model

Yvonne Stolz* and Joerg Baten**

This is the last working paper version before this study was submitted and accepted. Please cite as Stolz, Yvonne and Baten, Jörg (2012). "Brain Drain in the Age of Mass Migration: Does Relative Inequality Explain Migrant Selectivity?", *Explorations in Economic History* 49 (2012), pp. 205-20. Online: https://doi.org/10.1016/j.eeh.2012.01.001

Abstract

Emigration countries can suffer from a brain-drain. Do relative skill premia in source and destination countries matter for the brain-drain phenomenon? We explore human capital selectivity during the 1820s to 1900s period. In a sample of 52 source and five destination countries we find in fact brain drain effects determined by relative skill premia. Hence we confirm the Roy-Borjas model of migrant self-selection. Moreover, we find that countries like Germany and UK experienced a small positive effect, because the less educated migrated in larger numbers. We apply age heaping techniques to measure human capital selectivity of international migrants.

Keywords: International Migration, Labor Markets, Human Capital, Economic History
JEL codes: F22, J40, I21, N30

Acknowledgements: Seminar participants at the universities of Lund, Muenchen, and Tuebingen, and in particular Martin Biewen, Ray Cohn, Kerstin Enflo, Tim Hatton, Joachim Grammig, Willi Kohler, Joachim Voth and members of the Tuebingen economic history group provided valuable comments. Financial Support of the German Science Foundation (DFG), ESF GlobalEuroNet program and the EU HIPOD program is gratefully acknowledged, and data support of IPUMS, NAPP, Valeria Prayon, Matthias Blum, and Oliver de Marco. Eoin McLaughlin improved the language style.

*University of Tuebingen, yvonne.stolz@uni-tuebingen.de

**University of Tuebingen and CESifo, joerg.baten@uni-tuebingen.de

For countries with substantial emigration rates, brain drain is a core economic policy problem. This includes countries like Germany, which experiences both large-scale immigration and emigration nowadays. For example, the migration of Germans to the U.S. in the recent past has been discussed as brain drain, because the high U.S. skill premia attract a large number of highly skilled Germans. Chiswick (2005) summarized that often the "best and the brightest" would leave their home country to migrate to more promising labour markets.[1] Also in Africa, brain drain is perceived as an important issue (Docquier 2006). Although the health situation on the African continent is problematic, highly skilled African physicians leave and move in large numbers to the Western World because of higher returns to human capital. In a recent article in a leading medical journal, 'The Lancet', it was suggested that the recruitment of physicians from poor countries with high mortality ought to be treated as a criminal case because this would result in more people dying in the African source countries (Mills et al 2008). Consistent with those approaches, we define 'brain-drain' here as the phenomenon where, relative to the remaining population, a substantial number of more educated (numerate, literate) persons emigrate.

What determines the selectivity of migrants? Among other explanatory variables, relative skill premia or relative inequality have been stressed in the theory of self-selection. If the relative reward to skills is higher in the destination country than in the source country, we would expect highly skilled individuals to migrate, as Borjas (1987) formulated on the basis of Roy's self-selection model (Roy 1951). His views stimulated an excited debate, because those who came to the U.S. from high inequality countries such as Mexico were expected to be negatively selected (the well-educated Mexican might have stayed at home, as skill premia were high there). We will call this theoretical approach "Roy-Borjas model" in the following, as Borjas applied it to the process of migration.

The impact of skill premia, often proxied with inequality, on the selectivity of international migrants is still an open question of the literature. Brücker and Defoort (2006) find a positive correlation between inequality in the home country and educational selectivity of migrants in the OECD for the 1980-2000 period and develop a theoretical model that explains why more skilled people can cope with migration policy hurdles. Also, they find that inequality impacts positively on the human capital selectivity of migrants. Feliciano (2005) studies 32 immigrant groups in the US labor market and compares them with their source

[1] Of course, in today's world of skill-selective immigration policies, incentives in source countries sometimes also impact on acquiring a good education in order to have the choice to migrate, even if the more educated individual does not migrate in the end. Furthermore, international migrants send remittances to their home countries that also have an important developmental effect. For example in the Philippines, remittances made up just over 10 percent of national income in 2007. In Mexico and India, the figures have even been higher.

country's education and inequality level. She did not obtain results consistent with the Roy-Borjas model. However, Belot and Hatton (2009) find evidence for a modified Roy model for OECD immigration during the past decades.

We provide an analysis of a new and unique data set to this debate, as we can include international migrants from 52 source countries who went to five destinations in the Americas and Europe during the first era of globalization (1850-1910). The overall number of migrants included is 6.2 million. In the following we will aggregate them by decade, and by source and destination country pairs. As this period is the age of mass migration, our evidence provides a unique setting to investigate the question at hand, because migration flows were not yet mainly determined by immigration policies, which nowadays shape migrant selectivity significantly.[2] We include U.S. data until the 1900s, as the U.S. did not have strong immigration restrictions until 1919. Our Argentinean evidence covers only the migration until the decade of the 1880s (Timmer and Williamson 1996, Sanchez-Alonso 2008), as Argentina was the first to impose strong immigration restrictions starting mainly in the 1890s. Hence, the evidence studied here provides relatively undistorted evidence of migrant self-selection. We include not only major transatlantic destination countries, but also European immigration targets such as the UK. Finally, we also study one destination country which had actually more emigration than immigration: Norway had significant immigration from Sweden, Finland, Denmark, Germany, Iceland, but also Italy, UK, US and Russia. The major transatlantic destination countries are represented in our sample by the U.S., Canada and Argentina. This study is the first general assessment of migrant selectivity during this most crucial period of human migration history.

We apply the age heaping approach which captures basic numeracy skills by looking at the share of people who are able to report an exact age. In previous studies, this measure has always been found highly correlated with other education indicators (see, for example Crayen and Baten 2009). It will be explained in greater detail below. It allows the calculation of the difference between migrants' numeracy and numeracy of the source country population. We use this differential as the dependent variable and regress it on a set of explanatory variables. The paper is structured as follows: the next section briefly reviews the theory on human capital selectivity of international migrants and reviews results of earlier

[2] Germany, for example, attracted relatively low-skilled migrants during the 1960s and thereafter, because of the immigration policies at that time that aimed at providing unskilled labor for factory work, and the family unification allowances during the following period. Ireland on the other hand, attracted highly skilled labor in the recent decades which is partly due to its immigration policy, and partly due to large amounts of foreign direct investment before the economic crisis of 2009.

empirical studies. Then, we introduce the method, data and the model we estimate. In a next step, we discuss the results and make a number of robustness tests. We end with a conclusion.

2a. Theory: The Roy-Borjas model of relative skill premia

Economic theory implies that on a micro level, utility maximising individuals base their migration decisions on the benefits and costs of migration. Provided that the skill set a migrant incorporates is sufficiently applicable in the destination country, the expected gains from such a decision is the income gap between destination and home country multiplied with the probability of not being unemployed.[3] Migration costs comprise all the psychological, physical and material costs of the journey and subsequent settlement in a different environment. As migration always requires a certain amount of cash or "out-of-pocket-money" (Liebig et al 2004), and credit markets are normally imperfect, a poverty constraint exists, as the poorest often cannot pay for the migration cost. This is why, during the process of economic development, migration rises, when a country experiences initial economic growth because then the poverty constraint is less and less binding and more people can afford to migrate.

Migration costs increase with geographical and cultural distance, because travel costs and cultural costs (e.g. learning a language, religious differences etc) will be higher and the successful integration into the destination society might be more of a challenge. They decrease with growing diaspora communities in the target country, because friends and relatives living abroad might send remittances and provide valuable information, employment or other support for the newly arrived migrant.[4]

The impact of all these determinants on migration decisions is relatively well-documented in the literature. Hatton and Williamson (1998, 2004) have prominently shown that economic incentives played an important role throughout the mass migrations of the 19th century. What is less clear, however, is the question what determines migrant selectivity. Borjas (1987) developed a framework based on the Roy Model to approach the issue of migrant selectivity (Roy 1951). The basic model was originally formulated to explain individual self-selection into certain occupations and their impact on inequality, when an individual can chose between two possibilities. Given that the skills are sufficiently correlated among occupations, the individual will select into the occupation that provides the highest

[3] During the late 19th century, labor markets were not much regulated; hence obtaining a job at low wage was typically possible.

[4] Besides those variables, the importance of population growth to explain migration rates has been stressed, which translates on labor markets into a relative labor abundance in certain sectors which puts wages under pressure and can therefore make emigration more attractive to the affected individuals.

expected earnings. Borjas (1987) adapted the model to migration decisions. Here, the migrant selects himself into migration to a certain destination country, when his skill set will realize more income in the destination labor market than in the domestic one. An underlying assumption is that the skills can be applied in both countries and are sufficiently valued in both labor markets. A second condition is a market with sufficient information so that migrants are able to respond to those incentives. Is this realistic for the 19th century, our period of study? At least it is for the decision of some of the potential migrants in their source countries. Previous migrants often wrote letters informing their friends and relatives about the situation in the target country. While those letters were sometimes more optimistic than the real situation, they might have provided some information about the question whether unskilled or skilled workers were doing better, relative to the home country. Moreover, a large number of migrants reversed their decision if the benefits were not as large as expected and returned home.

To sum up, whether a person with a given skill level actually moves or not depends *ceteris paribus* on the relative skill premia of source and host country. Positive selection occurs when the destination displays a higher skill premium than the home country (see, for example German or African migration to the US in recent decades, or Russian Jews moving to 19th century U.S.). Negative selection occurs in the opposite case.

Belot and Hatton (2009) develop a variant of the Roy model to explain educational selectivity of migration flows into 29 OECD countries over the past decades. They also include immigration policy and poverty constraints. After controlling especially for poverty constraints of migration – as the poorest are not able to migrate – they obtain significant results for the inequality – selectivity link the Roy model proposes. Moreover, they find cultural and geographic distance to be very important.

Other empirical studies, in contrast, did not confirm the Roy-Borjas model. For example, Brücker and Defoort (2006) find a positive correlation between inequality in the home country and educational selectivity of migrants. Hence, the more unequal a country is, the better educated the emigrants will be. They argue that this is caused by higher abilities of the educated to jump over immigration restriction hurdles. Moreover, they find the same correlation for host country inequality. Feliciano (2005) finds no effect of income inequality on human capital selectivity for 32 immigrant groups in the US labor market, which also does not correspond with the Roy-Borjas model prediction. Hence, there exists no general agreement about the relationship between inequality and human capital selectivity of international migrants, yet.

Moreover, the issue has not been investigated from a historical and a broad international perspective until now. Wegge (2002) and Abramitzky et al. (2009) provide valuable studies on country cases and Cohn (2009) studies the early skill composition of mainly English, German, and Irish migrants to the United States 1820-1860 using the occupational composition of migrants as a proxy. Cohn makes clear that it was the migrants themselves, who declared the occupations. They sometimes tended to make exaggerated statements about their social and occupational status at home. In a review of Cohn's book, Kampfhoefner (2009) suggested to complement this approach with the age-heaping method. Mokyr (1983) pioneered these techniques for the Irish case (see also Ó Gráda (1986) for Baltic migrants to Dublin).

We extend those valuable historical studies by using the age-heaping indicator and by focusing on five destination countries and 52 source countries, offering systematic additional insights on this issue, taking a long-run, international approach.

3a. Other determinants of migrant selectivity

We expect transport costs and poverty constraints to play an important role. The log distance from the source country capital to the destination country capital multiplied with the decade-specific cost is included to proxy migration costs.[5] As the inhabitants of many poor countries and the poor within medium-income countries simply could not afford the transatlantic journey and many could not even afford migration within Europe, we need to control for poverty constraints. We subtracted GDP per capita from the maximum GDP per capita achieved in this period to obtain a measure of poverty.[6] As the poverty constraint might be less binding for a journey to a country which is closer, we interact logarithm of the distance with the deprivation measure to control for different intensity of this effect.

Another important component in the model is chain migration effects and remittances that earlier migrants might provide. Not only money is sent home, but also information about the destination country, which decreases the perceived risk of migration. Diaspora communities also provide valuable information and support in the form of money, employment, a shared language and identity, which makes the distance from home easier to bear. All these factors reduce the psychological and monetary costs of migration. Cohn (2009) argues that the friends-and-relatives effect decreases human capital selectivity of transatlantic

[5] The distance measure as well as data on colonial ties and common languages is taken from http://www.cepii.fr/anglaisgraph/bdd/distances.htm . On the decade-specific costs, see Sanchez-Alonso (2008).

[6] Where this was not available, we used imputations based on anthropometric values, see Baten 2006, and Baten and Blum 2010b.

migrants between 1820 and 1860. Especially during the last decades migrants were less positively selected from the underlying source country population. Around mid-century less skilled individuals could also afford the cost of passage and ever greater numbers of migrants wanted to escape the catastrophe of the hungry 1840s in many European countries. Mokyr (1983) confirms that the early migrants often reported occupations with high social status, but found that age heaping was significantly higher among Irish migrants than among the Irish population. While this is true for the whole pre-famine period, age-heaping on emigrant ships that arrive during the famine years is even higher. In some European countries the travel costs of the poor were even paid by the municipal communities which wanted to avoid the social transfers (von Hippel 1984, Bade 2008). This also contributed to less positively selected migrants. Our data set concentrates on the immigrants of those mid-century decades and thereafter, continuing in the U.S. case until 1910.

Apart from economic incentives, political, cultural and religious factors might also play a role. In German historiography, the democratic revolution attempt in 1848 and its aftermath generated an exodus of some highly educated individuals, who continued to play a role in American policies. In the regressions below, we test whether the German migration during mid-19th century displayed a different pattern because of this exogenous, political event. We also test whether the democracy situation in the destination country, relative to the source country, might have an impact on the selectivity of migrants.

Similarly, Eastern European migration was significantly shaped by religious factors. The Jewish minority experienced strong discrimination in the Russian Empire during this period, which reached its maximum in the pogrom waves of the 1880s. During the 1880s, the mass exodus of more than two million Russian Jews began. Already before, a migration stream of Jewish people started which was characterized by highly skilled individuals. This pronounced selectivity was not caused by economic incentives, but by political persecution. Therefore, we control in our regressions below – wherever possible – for such occasions.

Finally, we assess common language and colonial ties. Having to acquire a new language requires higher human capital of migrants than being able to use the mother tongue. We would hence expect pairs of countries with the same language to exhibit less positively selected individuals. On the other hand, advanced human capital can be more easily transferred between countries sharing the same language. This would suggest a positive effect on selectivity.

Colonial ties often show the same features. A common culture and institutions might make it easier for the migrant to adapt to the new environment. In the case of Britain, however, the type of colonial migrant might have been quite often government officials who went to the colonies to work in the administration or military. Their families might later return to Britain, in which case we would expect positive selectivity.

4a. Methodology

Age heaping is a method that uses the share of persons who report their exact age, as opposed to those who round erroneously, as an indicator for basic numeracy (Mokyr 1983, Crayen and Baten 2009a and 2009b). This indicator has been widely applied recently (A'Hearn, Baten and Crayen 2009, de Moor and van Zanden 2008, Clark 2007, Baten, Crayen and Voth 2008, Manzel and Baten 2009, Baten, Crayen and Manzel 2008, see also the applications in Humphries and Leunig 2009, Cinnirella 2008, O'Grada 2006). A'Hearn, Baten and Crayen (2009) have shown that within societies characterized by a lower level of human capital, the frequency of people stating their age erroneously is higher than in more developed societies. The tendency is to mention a convenient multiple of five instead of the exact age, which becomes evident in the frequency distribution of the age data. The ratio of the frequency of multiples of five in relation to the frequency of all mentioned numbers is defined as the Whipple Index.[7] The ABCC index employed below is a simple linear transformation of the Whipple index. It represents the estimated percentage share of the population who reported an exact age (A'Hearn, Baten and Crayen 2009).

The ABCC index correlates strongly with literacy rates, schooling and other human capital indicators, a relation which does not vary much across time and space and which is robust when applied to different types of data sources. Generally, the age heaping approach is considered a viable method to capture basic human capital in empirical studies. The great advantage of age heaping is the great variety of sources, where evidence can be drawn from. Further details are documented in Appendix B.

Interestingly, while some specialized studies have used the occupational structure and age heaping of migrants as indicators, the literacy of migrants was not used before. The reason might be the nature of literacy, which can be relatively easily achieved at higher ages, and which was demanded in some of the emigration countries such as the U.S. Unfortunately, literacy of immigrants at arrival was only assessed in the U.S. starting in 1899, when the U.S. public grew concerned with the educational status of recent mass immigration from Southern and Eastern Europe, and those lists are not available as individual data sets. Literacy was also

[7] The optimum is 100, i.e. an equal distribution of mentioned ages throughout the population, the extreme of 500 occurs, if everybody mentions a multiple of five only.

recorded in the censuses between 1850 and 1910, but the comparison between the literacy of immigrants in the U.S. and the population in the source country is difficult for a number of reasons.

Firstly, literacy in source countries was recorded using a number of different definitions. Some sources recorded literacy of the adult population, whereas the majority recorded those aged 15 and older, 10 and older or even six years and older.[8] Many statistics report just one number for the whole population which makes it impossible to calculate literacy of age groups or to obtain time series by birth cohorts.

Secondly, literacy of individuals coming from different linguistic backgrounds is always difficult to measure. Even if census takers were instructed to record literacy in any language and not only in the official language of the destination country, migrants from different language families could still have declared themselves illiterate when they were asked by census takers. We compared literacy and age heaping from the census data of the different migrant groups in the United States directly. Migrants with a Romanic-language background, namely Italy and Portugal, displayed average numeracy values. However, they had significantly lower literacy rates than one would expect according to their average numeracy.

Thirdly, although the vast majority arrived as young adults, a part of the migrants came as children and teenagers to the United States. Already for the mid-19th century, Cohn (2009) reports roughly one quarter arriving as children. When we look at the literacy of persons with migration background in the census some years later, we therefore have to be aware that many of them acquired literacy when they already lived in the United States. So the literacy performance is not only influenced by selective migration but also by age structure and schooling possibilities for migrants. To make things even more complicated, the U.S. was often the destination for migrants coming from countries with lower schooling (Eastern and Southern Europe), but also from countries with better schooling than the U.S., such as Sweden, Norway and so on. The children of those migrant families might have "lost" some of the schooling they would have obtained in their source countries if they had not

[8] In principle, we could calculate the literacy of U.S. citizens aged 10 and above, as the majority of source country literacy rates refers to that age range. However, the U.S. censuses of 1850 and 1860 did not display information on literacy of those aged less than 20.
We performed an exercise collecting all available literacy information, and regressed a dummy variable that controlled for children younger than 15 in the source on the literacy data base and obtained significant, negative coefficients for this variable. Hence, literacy skills during the nineteenth century were shaped by age structure and we can therefore not compare literacy data that contains children with data that does only contain adults, because the selectivity measure is sensitive to these distortions.

migrated. Therefore, there exist various biases of different directions which are difficult to quantify. For these reasons, the study of U.S. migrant selectivity based on literacy is too difficult at the present stage of knowledge. Fortunately, the age heaping techniques provides a feasible alternative to study this important issue.

A second methodological question was the measurement of skill premia. Although Borjas' original model looks at the standard deviation of wages, most of the literature on recent work on the Roy-Borjas model uses Gini coefficients of income distribution, because they are available for a large number of countries. The assumption of the literature is that wage variation and overall income Gini coefficients correlate. Belot and Hatton (2009) use skill premia directly measured in the wages for occupations that normally require some skills versus some that do not. We can summarize the previous literature saying that a broad mix of different inequality indicators was used. For the nineteenth century, skill premia are available for a number of countries. Since the 1920s, a large scale project has collected the wages of skilled and unskilled workers, especially in the building trades. This evidence has been recently been used in many studies (van Zanden 2009, Ljungberg 2008). In European countries of high inequality, a skilled artisan typically received about twice the wage of an unskilled labourer. In countries with lower inequality, the ratio was roughly 1.5. Since the 1930s the International Labor Organizations collected those wages by skill level for a large number of countries. For the remaining gaps, we used imputations of skill premia based on anthropometric inequality measures (Baten and Blum 2009).[9]

4b. Data

To measure human capital selectivity of migrants and compare them with the remaining population in the home country, it is necessary to measure both the human capital of migrants and of the population of the source country. For the migrants, we use data sets from the IPUMS and the North Atlantic Population Project that provide 100 percent census samples for the late 19th century for a number of countries, and smaller samples for other countries.[10] We only use information of individuals that are older than 23, because younger people are still able to recall their age more accurately. For the numeracy of source countries, we use published national censuses of a great number of countries that were originally compiled by Crayen and Baten (2009). Basic numeracy is acquired in the first decade of life. As it differed considerably between the different cohorts – and survivor biases and other biases turned out

[9] See notes to Table 3.

[10] See notes to Table 2.

insignificant in earlier studies – we can distinguish by age in each census. We obtain up to a maximum of five cohorts in each census (those aged 23-32, 33-42, …, 62-72). In the census data, the year of immigration is not noted. All previous migration studies found that the overwhelming majority migrated when they were around age 15-35, except for some children and a small number of older persons. Hence, we argue for the assumption that the period of migration decision must have been mostly two decades after birth. This has been counter-checked with lists created on ships, and we found the assumption justified. The ages 15-35 are by far the majority. Even more importantly, the numeracy by decade and country is almost exactly the same when looking at ship lists (with known time of migration) and census data. Comparing all passenger lists of ships arriving to New York between 1860 and 1895 the correlation of ABCC values by country and decade is 0.6 (p= 0.00, N=105).[11]

Geographically, we cover a wide range of source countries in Europe, Latin America, Asia, the Asian-Pacific and Africa to the US, the UK, Canada, Argentina and Norway as destination countries (Table 1). The global nature of our data set allows an in-depth analysis of international migration during the 19th century. The migration decades range from the 1820s up to the 1900s (Table 2). In this table, the average number of underlying observations is reported for each source country, decade, and destination country. For example, the average source contributed only 109 cases to Argentinean immigration in the 1830s, but 35,651 cases on average to U.S. immigration in the 1860s. Cases with less than 50 observations are excluded. The U.S. immigration before the 1880s is better documented than thereafter, because the NAPP project provided a 100% sample of the U.S. census in 1880, and smaller samples before and after.

Our dependent variable which measures human capital selectivity is constructed as a difference of the mean ABCC Index of migrants and the mean ABCC of the source country population. This measure is different from other migration selectivity studies which use the share of secondary and tertiary educated persons or years of schooling for recent decades (Belot and Hatton 2009). We took care to calculate the source country numeracy as a weighted share of stayers and migrants if the migration rates reach a substantial number, as

[11] We included all ship lists which were provided by the transcriber's guild (New York arrivals: http://www.immigrantships.net/nycarrivals1_6.html). Unfortunately, the number of observations is much smaller than in the case of census data – only some 300,000 compared to 6.2. million that we study here based on the census data -- hence we did not perform the same analysis with the ship lists. The advantage of ship list evidence is the possibility to determine the human capital status (and age) directly at arrival. One disadvantage is that it includes temporary migrants or travellers who returned home after a few months, but still the comparison to census data provides valuable insights. We thank Oliver de Marco for his immense contribution on this point.

during the time before the migration decision, the migrants' human capital still was part of the source country environment.[12]

One might argue that a bias might arise, if the census taking process in home and target country are different or the states are differently institutionalized and therefore ask their citizens with a different frequency for their ages. However, Crayen and Baten (2009) have shown that number of previous censuses taken as a proxy for institutionalized state-authority does not have a significant impact on the outcomes of the ABCC Index.

Another possible concern relates to the numeracy of the migrants, which is based on questions posed years after migration and therefore the migrant could have acquired some skills in the destination country. However, as mentioned above, we counter-checked our results with a sample of migrants that were obtained from ship lists, directly after arrival in the destination countries, the correlation was very close.

5. Results

5a. How did migrant selectivity develop during this period?

We first take a closer look at our dependent variable, which is defined as the numeracy of migrants minus the numeracy in the source country (both in percent). The average numeracy during this period in all source countries was 90 percent, in the destination countries 89 percent, hence almost equal. The average numeracy of migrants was 87 percent (arithmetic mean by source country). The weighted mean (weighted by migrant numbers) is quite similar, namely 86 percent numeracy. Hence there was no numeracy brain drain on average, but rather a mathematical brain gain for the source countries, because migrants who left in the 19th and early 20th century were slightly less numerate than the remaining population. But the difference is small. It is more interesting to look at the variation of brain drain and brain gain

[12] We used the migration numbers in Ferenczi and Willcox to identify the countries in which the migration rate exceeded one percent per decade to a given target country (in most cases, there was only one target country with such substantial migration). For the periods before 1870, we used the stock of migrants in the target countries, and compared overlapping numbers between Ferenczi and Willcox and census data in order to make sure that the differences in counting (Ferenczi and Willcox focus on migration statistics, hence an Irish migrant to Canada might have finally gone to the U.S; the census stock excludes those who died between migration and census taking). But the correspondence between both sources was quite good. For example, for the 1860s Ferenczi and Willcox list some 700,000 migrants from the UK (incl. Ireland) to the U.S., whereas the stock in the 1880s that we estimated to have migrated during the 1860s was 660,000.
We then calculated the weighted average of numeracy of stayers and migrants. Only for very few cases we had to assume similar values to the ones of other migrants (for example, we assumed that Spanish migration to Brazil in the 1880s was similar to the one to Argentina in the 1880s etc.).

between countries and over time and to study the determinants. In the following, we will take a look at some prominent examples of emigrant countries sending migrants to the US and UK. We arrange all numeracy values by migration decade.

The largest migrant flows to the United States in this period came from Germany and Ireland. Those migrants were mainly negatively selected for the early cohorts of our sample (Figure 1).[13] The German 1848 revolution does not show positive selectivity effects in our sample. We find actually 6-13 percent less numeracy among those migrating during the 1820s-1850s. Irish migrants display a strong negative selectivity, perhaps due to the Great Famine years, when remittances sent over by previous migrants were also used by the less educated to leave the country. Those who mainly migrated in the "hungry 1840s" display a value that is 20 percent lower than those, who stayed in Ireland. Over time, this negative selectivity diminishes and eventually dissolves completely for the migration cohorts 1880-1900.[14]

Among the "new immigration areas" in Eastern and Southern Europe – and the middle group of Swedish migration – the development is quite different (Figure 2). The Swedish and Italians show a very modest negative selectivity over the whole period with no major changes. In contrast, the Russian immigrants initially are very positively selected. The earliest cohorts migrating in the 1840s are more than 20% more numerate than their compatriots staying at home.[15] This is partly due to the fact that large shares of Russian immigrants were Jews, who have a reputation for better education than the overall population. Additionally, the high costs of migration from Eastern Europe translated in highly skilled first-wave migrants. Afterwards, there are probably strong "friends and relatives"-effects at work, probably also supported with remittances, as illustrated by the fact that the strong positive selectivity of the first decades decreases among the later cohorts. We should note though that the first decades of Russian migration were characterised by small absolute numbers.

Looking at another world region, we find immigration from Latin American countries positively selected. The absolute numbers here are small, which causes some volatility in the series (Figure 3). High migration costs could have caused brain drain for countries like Brazil, Peru and Chile to the US. The situation is less clear for Mexican immigrants, who had lower

[13] We consider Ireland separately, although it was part of the British Empire, because the characteristics of Irish migrants were different.

[14] Except for the small dip in German selectivity, which might have been caused by the economic crisis of the early 1890s, initiated by the Baring crisis.

[15] The immigration cohort of the 1830s would have been even more positively selected, but we removed it from the figure due to quite small sample size, in order not to provide an inadequate impression. Thanks to Ray Cohn for his important comment on this.

migration costs due to the geographic proximity. The early Mexican migrants tend to be equally or slightly negatively selected in terms of human capital in comparison to the home country population.

We cannot run through all five immigration countries, but as a second example, the English one is a particularly interesting case (Figure 4). Here, immigration is predominantly Irish in the first cohorts. These individuals are on average slightly positively selected (between 0 and 5 percent). Therefore, Ireland experiences some brain drain to England, but a brain gain migration to the US. Also, Poland and Russia, and to a lesser extent Canada suffer from brain drain effects due to migration to England. France and Germany, in contrast, did not experience brain drain with their modest migration flows to England.

In sum, although migrants are on average slightly negative selected, the variation between countries is large. Especially during the mid-19th century waves of migration, some of the main source countries display negative migrant selectivity partly caused by payments of source country government institutions who wanted to send away the poorest, and partly financed by remittances of earlier migrants (this was especially important for the Irish migration, see Cohn 2009). In contrast, Eastern European migrants are quite positively selected. Part of this migration is shaped by religious determinants. The Jewish minority experienced strong discrimination in the Russian Empire during the 19th century, which culminated in the persecution of the last decades of this century. The exodus consisted of individuals with much higher human capital. Economic incentives might have played a minor role in this case because skill premia in Russia were large.

5b. What determines migrant selectivity?

The base-line regressions are displayed in Table 3. We control with dummy variables for unobserved source country effects, destination country effects and time effects. This is important, because the relative skill premium argument of Borjas is a ceteris paribus argument: if the incomes in India are much lower, but the relative skill premium much higher than in the UK, we would still not expect that a large number of UK citizens would migrate (although those few who migrated might have been highly paid technicians and government officials). More generally, we employ destination and source country fixed effects to capture country specific political and socio-cultural characteristics and the income situation in destination and source countries as well.

As a result, *relative skill premia* seem to play a consistent role in determining migrant selectivity. The coefficient of this variable is positive and has the expected sign in all five specifications. In the first regression, we include the Russian emigration, although it might have been largely determined by religious factors, as explained in the previous section. In the second to fifth column, it is excluded and our results prove to be robust. In column 3, we tested a fixed effects model in order to control for unobserved heterogeneity (which is otherwise controlled with country dummies). The coefficient for relative skill premia is robust also in this specification. However, the Hausman test indicates that the random effects model applied in columns 1, 2, 4 and 5 should be preferred (Prob>chi2 = 0.9323). In column 4, we assess whether the skill premia matter only jointly with the friends and relatives effect, or poverty constraints, which is not the case.

Is the coefficient of relative skill premia economically meaningful? One measure for economic significance is to consider the effects of one standard deviation of the explanatory variable. If we multiply the standard deviation of skill premia (0.12, see Table 4) with its coefficient (11.79, col. 1 in Table 3), we obtain 1.47. This is roughly 18% of the standard deviation of the dependent variable (standard deviation: 8.16). Hence the importance is modest: it is neither small nor very large. If we do the same with the coefficient of the IV regression below (Table 6, column 5: 20.90), we obtain 2.60, which is around 32% of the standard deviation of the dependent variable. This is a substantial share, indicating economic significance.

The other variables had much less consistent effects. *The friends and relatives effect* is measured as the share of migrants coming from a specific source country and migrating to a destination country in the previous decade. It has always the expected negative sign, but is not statistically significant. *Poverty constraints* are calculated as the deviation of Log GDP (in a given country and decade) from the maximum Log GDP among all the countries during the period under observation.[16] The coefficient of this variable is sometimes significant (with an unexpected negative sign), and the same applies to distance. The poverty constraint can also be interpreted as an indicator of the lack of human capital in the countries of origin, which might explain its negative sign.[17]

We use several distance measures, like raw distance between the most populated locales in different countries, or a time variant measure of distance costs.[18] The latter measure of distance is more intuitive, as it can be considered as an estimate of “economic distance”.

[16] We use Maddison (2001), where not available, we imputed with anthropometric estimates (Baten and Blum 2009)

[17] Moreover, there might have been counter-acting forces, such as recruitment of service personnel, railway workers, “kulis” and other less skilled persons from poor countries such as China and Mexico, rather than the effect that only the rich could migrate from those poor countries.

[18] We took the passenger cost estimates by Sanchez-Alonso (2008), and calculated the cost for distance unit for each decade. This is then multiplied with actual distances. (distance measures from http://www.cepii.fr/)

The strong decline of transport cost with the arrival of the steamship innovation features prominently here. Nevertheless, this variable turns out to be insignificant (so, too, does the raw distance measure).

We also include an *interaction term between economic distance and poverty constraint*. It has the expected positive sign, but is insignificant. The poverty constraint is also insignificant, when included without interaction term in column 3. In a similar vein, we tested the *interaction between poverty constraint and migrant stock*, because we could have imagined that as poverty falls over time, the effect of migrant stock might become less negative. Hence, the expectation here would be a negative coefficient of this interaction term. However, it turns out to be insignificant (and positive) in column 2.

We also test for *relative democracy*, because one might expect that the more educated were attracted by higher democracy values in the destination country, relative to the source country. This is based on the estimates of democracy produced by the Polity IV project.[19] However, this variable turns out to be insignificant. The politically motivated migration might have been too small in number during the 19th century, and it was probably not sufficiently restricted to the educated strata. Finally, we also tested *common language* and *colonial relationships*, and found occasional positive effects (compare also the following Tables). In the case of the latter variable, this might have been caused by re-migration from colonial officials' families or similar special factors of the colonial administration. Common language might have been more useful for the more educated who had a comparative advantage with words and skills, rather than with brawn. Finally, in column 5 of Table 3 we test for a potential effect of *civil war* in the country of origin, which turns out to generate negative, but insignificant selectivity. Civil war countries seem to have been left predominantly by the less educated and poorer people during the period observed.

As a robustness test we omit in Table 5 some of the largest migrant groups, namely, the Germans, Irish and English. The results do not differ very much. If the Irish are omitted, the friends-and-relatives effects turns small and insignificant, this is neatly consistent with the literature that argued that the remittance effect was particularly important for Irish migration (Cohn 2009).

In the regressions of table 3 and 5, we consider each source country-destination country pair as one unit of decision making, only making sure with a minimum of 50

[19] Marshall, Monty G., and Jaggers, K.(2008): Polity IV Project: data set. http://www.systemicpeace.org/polity/polity4.htm#top

observations that the degree of measurement error is limited. An alternative is to weigh the observations with the number of migrants underlying each unit in a WLS regression which leads to more efficient estimates. Conventionally this is done with the square root of the number of underlying observations. The results in Table 6, column 1 and 2, are consistent with previous estimates. One potential disadvantage of weighted regressions is that a few source countries account for the majority of migrants; hence they receive most of the weight in the estimates.

We also queried whether the difference of migrant numeracy and source country numeracy might depend on the level of source country numeracy. Those coming from high education background might have been more likely to be negatively selected, even if we have seen many counter-examples in the Figures discussed above. We therefore include a term "ABCC level source country", which indeed turns significant but did not change the main results (Table 6, column 3). In a similar exercise to evaluate the properties of the dependent variable, we included only those in which the source country numeracy deviates from the optimum of 100 percent (Table 6, column 4). This removes some 35 cases, relative to column 2, but again the coefficients do not change.

A comprehensive test of time series properties indicated that the main series as well as the residuals do not display unit root problems. The Fisher test for unbalanced panels, as well as the Hadri-LM-test for the three largest source countries in a balanced panel, are calculated and suggest that our series do not suffer from unit root problems. [20]

Macro-economic analyses are under the permanent suspicion that endogeneity might play a role. A shock in the dependent variable might also lead to a significant change in one of the explanatory variables, and the direction of causality might be reversed. We are particularly interested in the question here whether the Roy-Borjas variable of relative skill premia could be endogenous. Could there theoretically be a mechanism of reverse causation, namely that migrant selectivity impacts on relative skill premia? In the long run, a massive exodus of a large share of a highly selective population could exert an influence on skill premia, if, for example, a large share of the unskilled workers leave. Then skill premia should ceteris paribus decline, if labor markets are functioning sufficiently. But this is less likely in the short run, from decade to decade, as we assess it here. The requirement of a large share of the population leaving is not what happened in most countries, where emigration rates were

[20] Next we fit a feasible generalized least squares model to assess the possibility of autocorrelation and the robustness of the results under this situation (Appendix Table 1, available from the authors). The autocorrelation term can be rejected with the 5 percent level of significance. However, even when we assume autocorrelation, the relative skill premia variable remains significant.

normally below 5 percent. Exceptions are Ireland where in some decades more than 10 percent left, and Italy right before WWI (Hatton and Williamson 1998).

Hence, endogeneity is only possible if the degree of selectivity and the migrant share of the population are large enough, so it depends on scale. In such cases, instrumental variables that make the scale criterion less likely to apply can help. Therefore, we have used an instrumental variable that considers the relative skill premia of world regions, such as Western Europe, Eastern Europe, and East Asia and so on. For example, one could imagine that selective Irish migration might impact on skill premia in Ireland, but much less so in all of Western Europe, because the numbers of Irish migrants were small relative to the large population of this world region. We calculated the inequality measure for all the world regions from which our migrants came and by decade. While the largest share of country-decade observations came from Western Europe (66%), also Eastern Europe contributed 11%, Latin America 9%, North America 8%, East Asia 5%, and very few observations stemmed from the Middle East and Southeast Asia. Hence, we have a sufficient variation by world region.

A second instrument comes from the political sphere. There was one major political effort to reduce inequality, which is closely related to skill premia. This effort was called the Bismarckian social insurance laws. The initiator was a conservative German politician who was concerned about the success of socialist movements in his country, and improving the living standard and security of the working masses seemed a sensible strategy to reduce inequality and hence the motivation of workers' associations. The social insurance laws consisted of sickness and injury insurances and the poorest that had lived in the strongest difficulties before benefited the most from this when they became sick. After Germany and Austria-Hungary started in the 1880s, other countries followed during the decades thereafter, and still others waited until the mid-20th century with those efforts (Flora 1982, Cutler 2002). This might have had an effect on migrant selectivity, but mainly via the potentially endogenous variable, skill premia and inequality, as required by instrumental variable analysis. The results of our instrumental variable regressions are shown in Table 6, Column 5, again confirming earlier results.

Finally, another approach to deal with potential endogeneity is to perform Arellano-Bond regressions, in which a large number of instrumental variables are generated from lagged first difference values of the dependent variable. Arellano and Bond (1991) suggested a consistent GMM estimator for this model in the presence of autoregression effects. This estimator is particularly strong if there are only a modest number of time observations, and a large number of cross-sectional observations, as is given in our sample. Arellano-Bond estimates have a reputation of being a very tough test of the robustness if endogeneity problems could be imagined. Their reliance on first differences also eliminates concerns about unit root problems. Again, the relative skill premia coefficients turn out robust (Table 7).

To conclude, a wide range of econometric techniques suggests that the relative skill premia had an effect on migrant selectivity as measured by relative numeracy with the age heaping method. There is some evidence -- although more limited -- on friends-and relatives-effects, colonial relationships and common language, whereas counteracting forces might have rendered the economic distance and democracy effects mostly insignificant.

Conclusion

In this study, we assess the selectivity of migrants in the era of mass migration. We focus not only on the main transatlantic migration destinations, but also on two European destination countries, the UK and Norway. Not less than 52 source countries could be included, and the underlying data set is based on 6.2 million individual migrants.

The main model tested here is the Roy-Borjas model, in which the selectivity of migrants is determined by the relative skill premia in destination and source countries. We confirm the influence of those economic migration incentives after controlling for a large number of other variables such as “friends and relatives effects”, poverty constraints, economic distance, relative democracy, common language and colonial relationships. This study has been the first general assessment of migrant selectivity during this most crucial period of human migration history, using large samples that included a variety of different source and destination countries.

It is crucial to understand the brain-drain processes between source and destination countries, because the stock of human capital determines future growth capabilities. Brain drain effects have not been systematically studied for the era of mass migration of the mid-to-late 19th century with large international samples before. In the case of mid-19th century mass migration history, there were also some arithmetic brain gains for the source countries, because those who left Scandinavia or central Europe around mid-century were often less numerate than the remaining population. For example, there could have been positive growth effects on Germany or Scandinavian countries, because the average numeracy must have increased due to migration. This process was reinforced by remittances. In contrast, Eastern Europe lost a large number of the numerate population, and the migration effects might have been *ceteris paribus* negative for Eastern Europe. Clearly, a large number of other factors were also at work, hence these effects should not be seen in isolation. But understanding migrant selectivity helps to identify an additional and important variable in the global long-term growth record.

References

Abramitzky, R., Boustan, L.P., Eriksson, K. 2009. “Measuring Selectivity and Returns in the age of Mass Migration”. NBER Working Paper 15684.

Arellano, M., and S. Bond. 1991. “Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations.” *Review of Economic Studies* 58: 277-297.

Bade, K.J. (ed.). 2008. *Enzyklopädie Migration in Europa: vom 17. Jh. bis zur Gegenwart*. Stuttgart.

A’Hearn, B., Baten, J., Crayen, D. 2009. “Quantifying Quantitative Literacy: Age Heaping and the History of Human Capital.” *The Journal of Economic History* 69/3, pp. 783-808.

Baten, J., Crayen, D., Manzel, D. 2008. “Numerical Abilities and Numerical Discipline in Northern and Western Germany, 16th to 18th Centuries.” *Jahrbuch fuer Wirtschaftsgeschichte* 2008/2, pp. 217-229.

Baten, J., Foldvari, P., Leeuwen, B.v., Luiten van Zanden, J. 2009. “World Income Inequality 1820-2000.” Working Paper University of Tuebingen, Dept. Economic History No. 87.

Baten, J., Crayen, D., Voth, J. 2008. “Poor, Hungry and Stupid: Numeracy and the Impact of High Food Prices in Industrializing Britain, 1780-1850.” Working Paper Universitat Pompeu Fabra. Departamento de Economía y Empresa , Nº. 1120, 2008

Belot, M.V.K., Hatton, T.J. 2009. “Immigrant Selection in the OECD.” Center for Economic Policy Research. Discussion Paper No. 571.

Blum, M. 2009. “The Influence of Inequality on the Standard of Living: Worldwide Evidence from 1810 to 1980. University of Tuebingen. Working Paper.

Borjas, G.J. 1987. “Self-Selection and the Earnings of Immigrants.” *The American Economic Review* 77/4, pp. 531-553.

Brücker H, Defoort, C. 2006. „The Self-Selection of International Migrants Reconsidered: Theory and New Evidence.” IZA Discussion Paper Series. IZA DP 2052.

Chiswick, B.R. 2005. „High Skilled Immigration in the International Arena.” IZA Discussion Paper Series. IZA DP 1782.

Cinnirella, F. 2008. “Optimists or pessimists? A reconsideration of nutritional status in Britain, 1740–1865.” *European Review of Economic History* 2008/12, pp. 325-354.

Clark, G. 2007. *A Farewell to Alms: A Brief Economic History of the World.* Princeton University Press.

Cohn, R.L. 2009. *Mass Migration under Sail: European Immigration to the Antebellum United States.* Cambridge University Press. Cambridge.

Crayen, D., Baten, J. 2009. “Global Trends in Numeracy 1820–1949 and its Implications for Long-term Growth.” *Explorations in Economic History* 47/1, pp. 82-99.

Crayen, D., Baten, J. 2010. “New Evidence and New Methods to Measure Human Capital Inequality before and during the Industrial Revolution: France and the US in the Seventeenth to Nineteenth centuries.” *Economic History Review,* forthcoming. Published Online: Jul 29 2009 10:50AM. DOI: 10.1111/j.1468-0289.2009.00499.x

Cutler, David M. and Richard Johnson 2002, “The Birth and Growth of the Social Insurance State: Explaining Old Age and Medical Insurance Across Countries.” WP Harvard U /Kansas Fed.

De Moor, T. and Van Zanden, J.-L. 2008. “Uit fouten kun je leren. Een kritische benadering van de mogelijkheden van 'leeftijdstapelen' voor sociaal-economisch historisch onderzoek naar gecijferdheid in het pre-industriële Vlaanderen en Nederland.” *Tijdschrift voor Economische en Sociale Geschiedenis* 5-4: 55-86.

Docquier, F. 2006. “Brain Drain and Inequality Across Nations.” IZA Discussion Paper Series. IZA DP 2440.

Feliciano, C. 2005. “Educational Selectivity in U.S. Immigration: How Do Immigrants Compare to those left behind?” *Demography* 42/1, pp. 131-152.

Flora, P. 1983. State, *Economy and Society in Western Europe: 1815-1975. A data handbook in two Volumes.* Frankfurt, M. Campus.

Hatton, T.J., Williamson, J.G. 1998. *The Age of Mass Migration: Causes and Economic Impact*. Oxford University Press. New York.

Hatton, T.J., Williamson, J.G. 2008. *Global Migration and the World Economy: two Centuries of Policy and Performance.* MIT Press. Michigan.

Hippel, W. von 1984. *Auswanderung aus Südwestdeutschland. Studien zur Württembergischen Auswanderung und Auswanderungspolitik im 18. und 19. Jahrhundert.* Stuttgart.

Humphries, J., Leunig, T. 2009. “Was Dick Whittington taller than those he left behind? Anthropometric measures, migration and the quality of life in early nineteenth century London?” *Explorations in Economic History* 46/1, pp. 120-131.

Kamphoefner, W. 2009. “Mass Migration under Sail: European Immigration to the Antebellum United States. By Raymond L. Cohn. Book Review.” *The Journal of Interdisciplinary History* 40/4, pp. 621-622.

Maddison, A. 2009. *The World Economy: a millennial Perspective*. OECD Publishing 2001.

Manzel, K., Baten, J. 2009. “Gender Equality and Inequality in Numeracy: The Case of Latin America and the Caribbean 1880–1949.” *Revista de Historia Económica/Journal of Iberian and Latin American Economic History* 2009, Second Series, 27, pp. 37-73.

Milanovic, B. 2009. “Global Inequality and Global Inequality Extraction Ratio: the story of the last two Centuries.” World Bank Policy Research Working Paper No. 5044.

Mills, E.J., Schabas, W.A., Volmink, J., Walker, R. , Ford, N., Katabira, E., Anema, A., Joffres, M., Cahn, P., Montaner, J. 2008. “Should active recruitment of health workers from sub-Saharan Africa be viewed as a crime?” *The Lancet* 371, pp. 685-88.

O’Grada, C. 1986. “Across the Briny Ocean: some thoughts on the Irish emigration to America 1800-1850.” *Migrations across time and nations: population mobility in historical contexts*. New York, pp. 79-94.

O’Grada, C. 2006. “Dublin Jewish Demography a Century Ago.” *The Economic and Social Review,* Vol. 37, No. 2, Summer/Autumn, 2006, pp. 123-147.

Pradhan, M., Sahn D.E., Younger, S.D 2002. “Decomposing World Health Inequality.” *Journal of Health Economics* 22/2, pp.271-293.

Rodriguez Galdo, M.X., Cordero Torrón, X. 2007. “Emigración e Mercado de Traballo. Espanois en Arxentina 1882-1926.” *Revista Galega de Economía*, 16. Numero extraordinario, pp. 93-116.

Roy, A. 1951. “Some thoughts on the distribution of earnings.” *Oxford Economic Papers* 3, pp. 135–46.

Sanchez Alonso, B. 2008. “The Other Europeans: Immigration into Latin America and the International Labour Market, 1870-1930” *Revista de Historia Económica/ Journal of Iberian and Latin American Studies, XXV, 3,* 2007, pp. 395-426

Timmer, A.S., Williamson, J.G. 1996. “Racism, Xenophobia or Markets? The Political Economy of Immigration Policy Prior to the Thirties.” NBER Working Paper W5867.

Wegge, S.A. 2002. “Occupational Self-selection of Nineteenth-Century German Emigrants: Evidence from the Principality of Hesse-Cassel.” *European Review of Economic History* 6 (3), pp. 365-394.

Tables and Figures

Table 1: Underlying number of cases by source country

Country	Cases	Country	Cases	Country	Cases
Ireland	1877232	Mexico	45828	Barbados	1841
Germany	1719228	Netherld.	42674	India	1208
UK	978433	Austria	32834	Iceland	1159
Canada	467201	France	29777	Uruguay	1050
Sweden	205227	Russia	26044	Greece	845
Norway	138013	Portugal	11362	Brazil	844
Belgium	101223	Luxembg	10902	Hong Kong	812
China	86092	Spain	9274	Turkey	812
Switz.ld	75371	Hungary	8589	Romania	751
Czech	60458	Finland	8021	Jamaica	670
Denmark	53816	Cuba	4683	Japan	548
Italy	51385	Australia	2229	Bermuda	430
US	47985	Chile	1978	Bolivia	244
Poland	46183				

Sources: see notes to Table 2.

Table 2: Average number of underlying cases for each decade, destination and source country, by decade and destination country

Destination	1820	1830	1840	1850	1860	1870	1880	1890	1900
Argentina		109	265	428	721	655	623		
Canada	197	10664	8723	7228	5220	4352	381	421	
Norway	527	878	1490	2511	2120	2002	1798	1655	
UK	1451	1208	1611	515	411	376			
US	915	13006	24900	32064	35651	30703	989	941	655

Notes: For example, 109 was the average number of cases of all source countries that provided migrants to Argentina in the 1830s.

Census evidence was available for Argentina (1869, 1895) – sample; Canada (1871, 1881-100%, 1901); Norway (1865, 1875, 1900); England (1851, 1881); US (1850, 1860, 1870, 1880-100%, 1890, 1900, 1910).

Sources: On the U.S. except 1880: Ruggles, Steven, Matthew Sobek, and Trent Alexander, et al. *Integrated Public Use Microdata Series: Version 3.0* [Machine-readable database]. Minneapolis, MN: Minnesota Population Center [producer and distributor], 2004. On Argentina : Somoza, J. and Lattes, A. (1967): Muestras de los dos primeros censos nacionales de población, 1869 y 1895. Documento de Trabajo No 46, Instituto T. Di Tella, CIS, Buenos Aires. On all other samples: North Atlantic Population Project and Minnesota Population Center. NAPP: Complete Count Microdata. NAPP Version 2.0 [computer files]. Minneapolis, MN: Minnesota Population Center [distributor], 2008. [http://www.nappdata.org];

Table 3: Regression of human capital selectivity (numeracy migrant in % - numeracy source country in %)

	(1)	(2)	(3)	(4)	(5)
Estimation method	RE	RE	FE	RE	RE
Source countries excluded	None	Russia	Russia	Russia	Russia
Relative skill premium dest - source	11.79***	11.11***	13.30***	7.97**	9.58**
	(0.0030)	(0.0045)	(0.00022)	(0.040)	(0.012)
Friends & relatives (Ln stock mig)	-0.43	-0.61	-0.46		-0.13
	(0.22)	(0.41)	(0.46)		(0.67)
Poverty constraint (max GDP - GDP)	-9.47*	-1.34	-8.05***		
	(0.075)	(0.70)	(0.0012)		
Ln distance	-2.43**	-1.29*			
	(0.022)	(0.99)			
Ln distance * poverty constraint	2.22				
	(0.21)				
Poverty constraint * Friends & rel.		0.20			
		(0.81)			
Relative democracy	-0.44	-0.37			
	(0.39)	(0.46)			
Common Language	2.94	4.20**			
	(0.18)	(0.050)			
Colonial relationship	1.01	1.20			
	(0.65)	(0.59)			
Civil war					-1.81
					(0.11)
Destination	Yes	Yes	Yes, FE	No	Yes
Source	Yes	Yes	Yes, FE	Yes	Yes
Time	Yes	Yes	No	No	No
Constant	-1.36	-0.77	1.66	-10.70***	-11.51**
	(0.82)	(0.89)	(0.39)	(0.0061)	(0.016)
Observations	303	291	297	376	300
R-squared (within)	0.16	0.20	0.09	0.11	0.20
R-squared (within)	0.59	0.58	0.04	0.54	0.56

P-values based on robust standard errors are included in brackets. Migration decades 1820s-1900s are included. Column 1, 2, 4 and 5 are estimated with random effects models (but country dummies included), col. 3 is based on fixed effects estimates.

Sources: on the migrant numeracy, see Table 2. Numeracy in the source countries are from Crayen and Baten (2009). Skill Premia are from Baten and Blum (2010a), Estimating Skill Premia with Anthropometric Indicators, see http://www.wiwi.uni-tuebingen.de/cms/fileadmin/Uploads/Schulung/Schulung5/Joerg/Baten_Blum_skpr100331a.pdf, last accessed March 31st, 2010. The stock of migrants was calculated with migrant data sets cited in the notes to Table 2. Poverty constraints are based on Maddison (2001), and for those countries for which values were lacking we used the imputations first done by Baten and Blum (2010b), see http://www.wiwi.uni-tuebingen.de/cms/fileadmin/Uploads/Schulung/Schulung5/Joerg/baten_blum_ht_100331a.pdf last accessed March 31st, 2010. The distance measure as well as data on colonial ties and common languages is taken from http://www.cepii.fr/anglaisgraph/bdd/distances.htm last accessed March 31st, 2010. The distance was then multiplied with the passenger cost estimates by Sanchez-Alonso (2008) to account for the decline in distance costs. Relative democracy data is from the Polity IV project, see Marshall, Monty G., and Jaggers, K.(2008): Polity IV Project: data set. http://www.systemicpeace.org/polity/polity4.htm#top last accessed March 31st, 2010. Civil War data is from the Correlates of War Project, see Singer, J. David and Melvin Small (1972): The Wages of War, 1816-1965: A Statistical Handbook. New York. Or see http://www.correlatesofwar.org last accessed March 31st, 2010.

Table 4: Descriptive statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
Migrant selectivity	303	-4.32	8.16	-27.91	52.00
Relative skill premia	303	-0.06	0.12	-0.42	0.33
Ln migrant stock	303	0.60	2.14	-4.98	4.61
Poverty constraint	303	0.72	0.27	0.00	1.72
Ln distance	303	3.01	1.07	0.48	4.60
Ln dist*pov. constr.	303	2.24	1.27	0.00	6.79
Relative democr.	303	1.34	3.50	-5.70	10.00
Common language	303	0.23	0.42	0.00	1.00
Colonial r'ship	303	0.15	0.36	0.00	1.00

Note: only the cases are included for which all explanatory variables (Table 3, Col 1) did not contain missing values. Sources: see Table 2 and 3.

Table 5: Robustness of human capital selectivity regression: excluding the largest source countries

	(1)	(2)	(3)
Source countries excluded	Germany	Ireland	UK
Relative skill premium dest - source	10.12***	9.76**	10.55***
	(0.0098)	(0.012)	(0.0061)
Friends & relatives (Ln stock mig)	-0.69*	-0.09	-0.49
	(0.099)	(0.78)	(0.23)
Poverty constraint (max LGDP - LGDP)	-3.42	-5.40	-5.99
	(0.52)	(0.25)	(0.22)
Ln distance	-2.24**	-0.93	-2.03*
	(0.047)	(0.36)	(0.057)
Ln distance * poverty constraint	1.36	1.19	1.74
	(0.47)	(0.46)	(0.32)
Common Language	5.07**	5.38***	3.40
	(0.033)	(0.0091)	(0.15)
Colonial relationship	1.16	-2.30	-0.01
	(0.64)	(0.37)	(1.00)
Destination	Yes	Yes	Yes
Source	Yes	Yes	Yes
Time	Yes	Yes	Yes
Constant	4.09	-2.94	3.55
	(0.56)	(0.64)	(0.59)
Observations	269	278	274
R-squared (within)	0.23	0.15	0.18
R-squared (overall)	0.59	0.60	0.58

P-values based on robust standard errors are included in brackets. Migration decades 1820s-1900s are included. Russia excluded. Sources: see Table 2 and 3.

Table 6: Regression of human capital selectivity, weighted by number of underlying observations, and IV estimation

	(1)	(2)	(3)	(4)	(5)
Estimation method	LSDV	LSDV	LSDV	LSDV	IV
	WLS	WLS	WLS	WLS	WLS
Included abcc range	All	All	All	<100	All
Relative skill premium dest - source	9.74**	9.45**	10.64***	10.71**	20.90**
	(0.017)	(0.019)	(0.00")	(0.012)	(0.044)
Friends & relatives (Ln stock mig)	-0.07	-0.07	0.14	-0.48	-0.17
	(0.79)	(0.79)	(0.58)	(0.12)	(0.46)
Poverty constraint (max GDP - GDP)	-0.93	-0.91	6.78*	-1.94	-8.28
	(0.82)	(0.83)	(0.085)	(0.71)	(0.24)
Ln distance	-2.56***	-2.65***	-1.22	-2.93***	-5.22***
	(0.00%)	(0.99)	(0.99)	(0.99)	(0.99)
Ln distance * poverty constraint	1.12	1.33	-0.98	1.43	3.23
	(0.42)	(0.34)	(0.47)	(0.40)	(0.11)
Relative democracy	0.18				
	(0.64)				
Common Language	-1.00	-0.99	-1.19	-0.37	-6.55***
	(0.48)	(0.49)	(0.38)	(0.81)	(0.000)
Colonial relationship	6.19***	6.12***	6.61***	6.50***	4.35***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.00=)
ABCC level source country			-0.73***		
			(0.000)		
Destination	Yes	Yes	Yes	Yes	Yes
Source	Yes	Yes	Yes	Yes	No
Time	Yes	Yes	Yes	Yes	No
Constant	5.54	2.05	70.03***	5.66	8.99**
	(0.25)	(0.71)	(0.000)	(0.34)	(0.014)
Observations	303	309	309	264	297
R-squared	0.64	0.64	0.71	0.65	0.16

P-values based on robust standard errors are included in brackets. Migration decades 1820s-1900s are included.. Russia excluded. Sources: see Table 2 and 3.

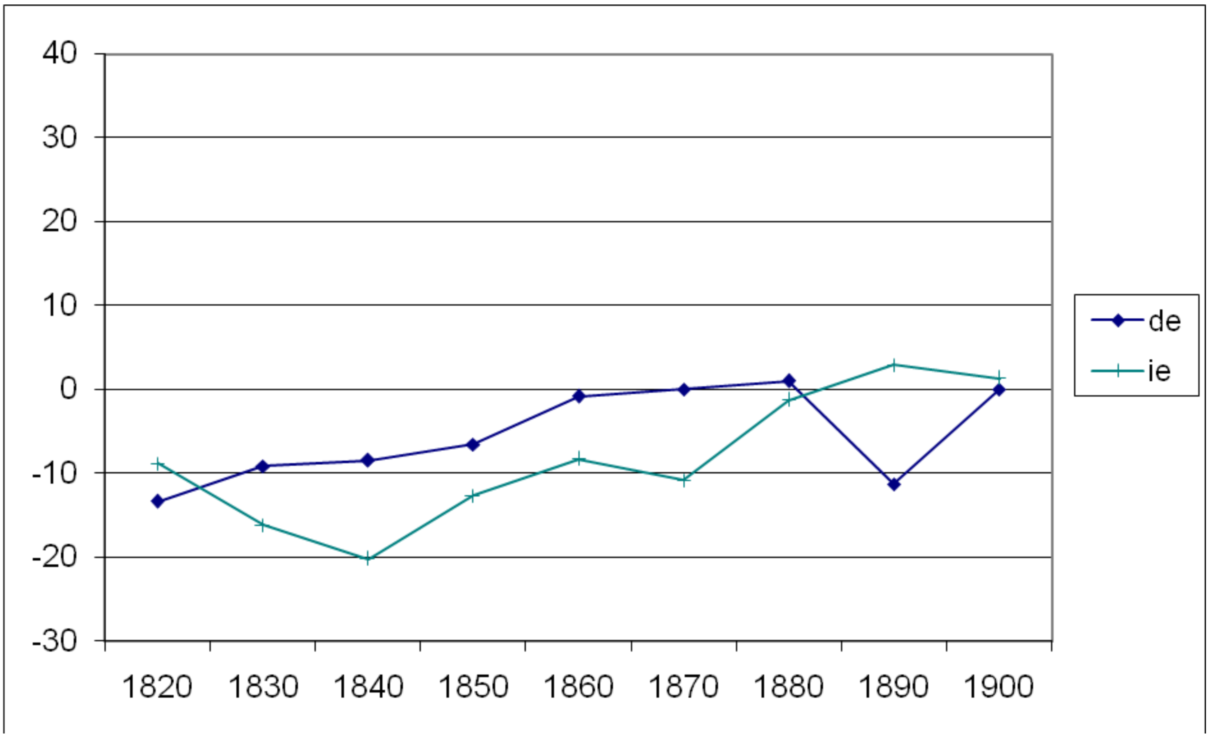
Instrumental variables: Dummy variable “Social Insurance reforms”, and relative skill premia by world region. Migration decades 1820s-1900s are included. Russia excluded. Tests of overidentifying restrictions (IV in col. 5): Sargan (score) chi2(1) = .02566 (p = 0.8727), hence passed.

Table 7: Arellano Bond dynamic panel regressions

	(1)	(2)
Relative skill premium dest - source	8.92***	7.47***
	(0.0019)	(0.0033)
Lagged migrant selectivity	0.21	0.08
	(0.11)	(0.52)
Friends & relatives (Ln stock mig)	0.17	-0.39
	(0.81)	(0.59)
Poverty constaint (maxLGDP-LGDP)		-12.35
		(0.22)
Log distance		-6.31*
		(0.098)
Ln distance * poverty constraint		2.39
		(0.46)
Constant	-3.50***	18.52*
	(0.000018)	(0.083)
Observations	228	226
No(instruments)	45	48
p-value of Wald chi2	0.002	0.000

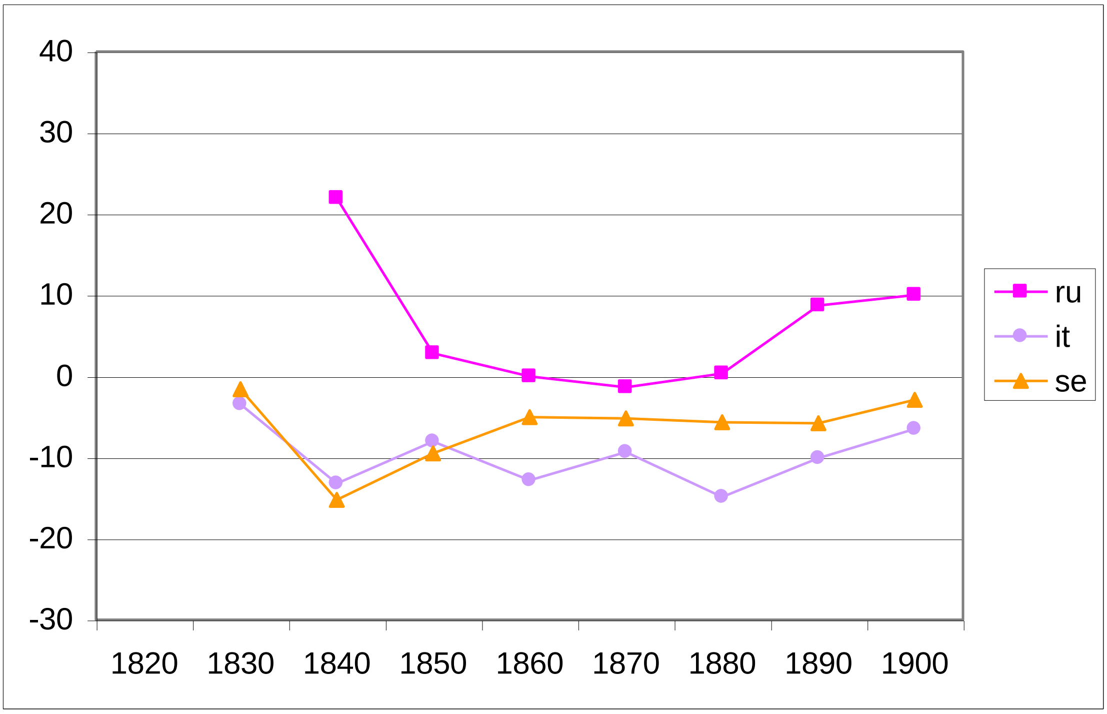
Migration decades 1820s-1900s are included. Russia excluded. We use the entire lag structure for instrumentation, i.e. starting from the (t-2) lag of the difference for the levels equation, and the (t-1) lag of the level for the difference equations. Arellano-Bond test for AR(2) in first differences. Prob > z: 0.22, hence passed. The Sargan test of overidentifying restrictions yielded a chi2 of 47.26 (Prob > chi2 = 0.22), hence passed. Sources: see Table 2 and 3.

Figure 1: Selectivity among migrants from Germany and Ireland ("old migration countries") to the U.S.



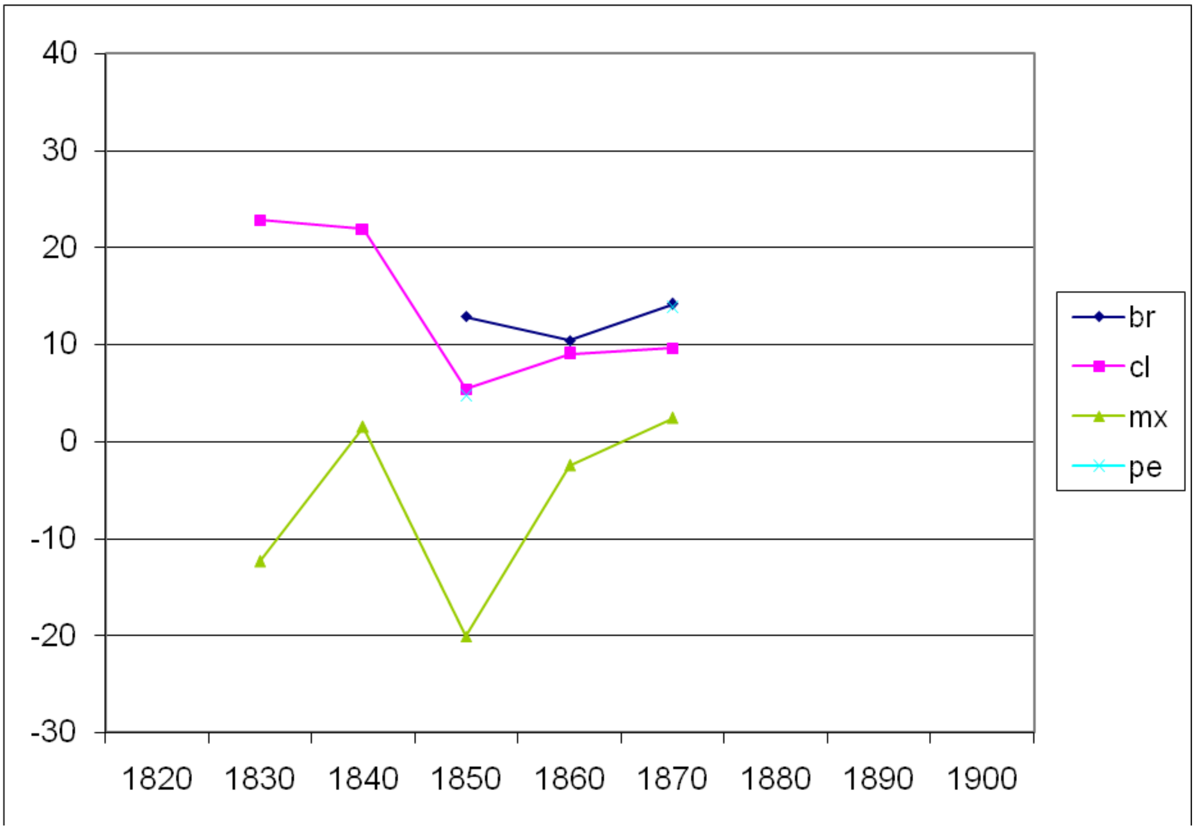
Sources: see Table 2.

Figure 2: Selectivity among migrants from middle and “new” migration countries to the U.S.



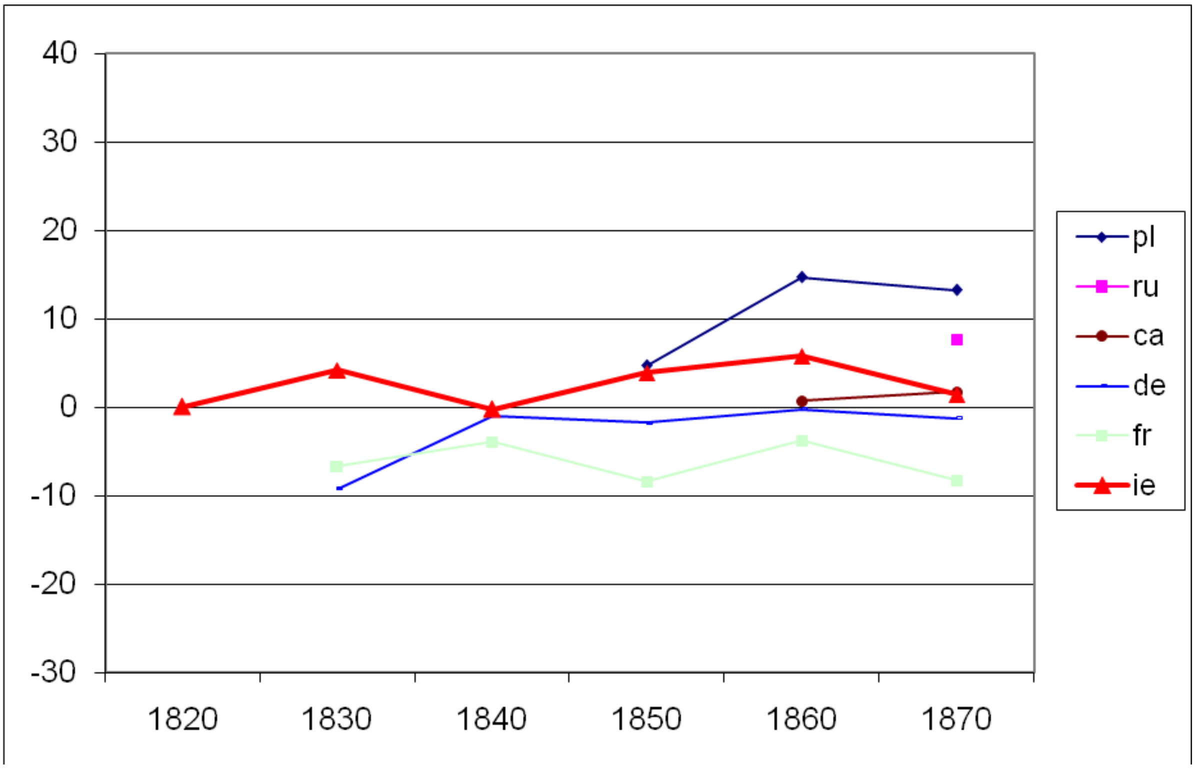
Sources: see Table 2.

Figure 3: Selectivity among Latin American migrants to the U.S.



Sources: see Table 2.

Figure 4: Selectivity among several migrant groups to England



Sources: see Table 2.

Not for publication: Appendix A: Autocorrelation. Appendix Table 1: Feasible GLS regressions, assuming an AR(1) process

	(1)
	GLS (AR1)
Relative skill premium dest - source	8.78**
	(0.023)
Friends & relatives (Ln stock mig)	-0.14
	(0.58)
Poverty constraint (maxLGDP-GDP)	-11.50**
	(0.050)
Ln distance	-3.43***
	(0.99)
Ln distance * poverty constr	4.15**
	(0.020)
Common Language	3.34**
	(0.044)
Colonial relationship	1.03
	(0.54)
Destination	YES
Source	YES
Time	YES
Constant	-1.71
	-0.74
Observations	297
Wald chi2(43)	260.49
p-value (Wald)	0.00
common AR(1) coefficient for all panels	0.082

P-values based are included in brackets. Migration decades 1820s-1900s are included. Russia excluded. Sources: see Table 2.

Not for publication: Appendix B: Methodology and basic concepts of age heaping (Internet Appendix)

We study numerical abilities in this article, which are an important component of overall human capital. In order to provide estimates of very basic components of numeracy, we will apply the age heaping methodology.[21] The idea is that in less developed countries of the past, only a certain share of the population was able to report the own age exactly when census-takers, army recruitment officers, or prison officials asked for it. The remaining population reported a rounded age, for example, 40, when they were in fact 39 or 41. In today's world of obligatory schooling, passports, universities, birth documents, and bureaucracy, it is hard to imagine that people did not know their exact age. But in early and less organized societies this was clearly different. The typical result is an age distribution with spikes at ages ending in a five or a zero and an underrepresentation of other ages, which does not reflect the true age distribution. There was also some heaping on multiples of two, which was quite widespread among children and teenagers and to a lesser extent among young adults in their twenties. This shows that most individuals actually knew their age as teenagers, but only in well-educated societies were they able to remember or calculate their exact age again later in life.[22]

To give an example of rounding on multiples of five, the census of Mexico City 1790 reports 410 people aged 40, but only 42 aged 41. This was clearly caused by age heaping. Apolant (1975, p. 333) gives individual examples of age misreporting: Joseph Milan, who appeared in February 1747 as a witness in an Uruguayan court, should have been 48 years old, according to one judicial record. However, in the same year, but in another judicial record, he declares his age to be '45 years'. Demographers see this age misreporting as a problem when calculating life expectancies and other population statistics. But exactly this

[21] For more detailed surveys on the age heaping methodology see A'Hearn, Baten and Crayen (2009).

[22] At higher ages, this heaping pattern is mostly negligible, but interestingly somewhat stronger among populations who are numerate enough not to round on multiples of five.

misreporting enables us to approximate numerical abilities of historical populations. The ratio between the preferred ages and the others can be calculated by using several indices, one of them being the Whipple index.[23] To calculate the Whipple index of age heaping, the number of persons reporting a rounded age ending with 0 or 5 is divided by the total number of people, and this is subsequently multiplied by 500. Thus, the index measures the proportion of people who state an age ending in a five or zero, assuming that each terminal digit should appear with the same frequency in the 'true' age distribution.[24]

$$(1)\ Wh = \left(\frac{\sum (Age25 + Age30 + ...Age60)}{1/5 \times \sum Age23 + Age24 + Age25 + .. + Age62)} \right) \times 100$$

For an easier interpretation, A'Hearn, Baten, and Crayen (2009) suggested another index, which we call the ABCC index.[25] It is a simple linear transformation of the Whipple index and yields an estimate of the share of individuals who correctly report their age:

$$(2)\ ABCC = \left(1 - \frac{(Wh - 100)}{400}\right) \times 100 \text{ if } Wh \geq 100 \text{; else } ABCC = 100 \text{.}$$

The share of persons able to report an exact age turns out to be highly correlated with other measures of human capital, like literacy and schooling, both across countries, individuals, and over time (Bachi 1951, Myers 1954, Mokyr 1983, A'Hearn, Baten, and Crayen 2009). A'Hearn, Baten, and Crayen (2009) found that the relationship between illiteracy and age heaping for less developed countries (LDCs) after 1950 is very close. They calculated age heaping and illiteracy for not less than 270,000 individuals who were

[23] A'Hearn, Baten and Crayen (2009) found that this index is the only one that fulfils the desired properties of scale independence (a linear response to the degree of heaping), and that it ranks samples with different degrees of heaping reliably.

[24] A value of 500 means an age distribution with ages ending only on multiples of five, whereas 100 indicates no heaping patterns on multiples of five, that is exactly 20 percent of the population reported an age ending in a multiple of five.

[25] The name results from the initials of the authors' last names plus Greg Clark's, who suggested this in a comment on their paper. Whipple indexes below 100 are normally caused by random variation of birth rates in the 20th century rich countries. They are not carrying important information, hence normally set to 100 in the ABCC index.

organized by 416 regions, ranging from Latin America to Oceania.[26] The correlation coefficient with illiteracy was as high as 0.7. The correlation with the PISA results for numerical skills was even as high as 0.85, hence the Whipple index is more strongly correlated with numerical skills. They also used a large U.S. census sample to perform a very detailed analysis of this relationship. They subdivided by race, gender, high and low educational status, and other criteria. In each case, they obtained a statistically significant relationship. Remarkable is also the fact that the coefficients are relatively stable between samples, i.e., a unit change in age heaping is associated with similar changes in literacy across the various tests. The results are not only valid for the U.S.: In any country with substantial age heaping that has been studied so far, the correlation was both statistically and economically significant.

In order to assess the robustness of those U.S. census results and the similar conclusions drawn from late 20th century LDCs, A'Hearn, Baten, and Crayen (2009) also assessed age heaping and literacy in 16 different European countries between the Middle Ages and the early 19th century. Again, they found a positive correlation between age heaping and literacy, although the relationship was somewhat weaker than for the 19th or 20th century data. It is likely that the unavoidable measurement error when using early modern data caused the lower statistical significance.

Age heaping has also been compared to other human capital indicators, for example, primary schooling rates. The widest geographical sample studied so far was created by Crayen and Baten (2009), who were able to include 70 countries for which both age heaping and schooling data (as well as other explanatory variables) were available. They found in a series of cross-sections between the 1880s and 1940s that primary schooling and age heaping were closely correlated, with R-squares between 0.55 and 0.76 (including other control variables; see below). Again, the coefficients were relatively stable over time. This large

[26] See A'Hearn, Baten and Crayen (2009), Appendix available from the authors.

sample also allowed the examination of various other potential determinants of age heaping. To assess whether the degree of bureaucracy, birth registration, and government interaction with citizens are likely to influence the knowledge of one's exact age, independently of personal education, the authors used the number of censuses performed for each individual country for the period under study as an explanatory variable for their age heaping measure. Except for countries with a very long history of census-taking, all variations of this variable turned out insignificant, which would suggest that an independent bureaucracy effect was rather weak. In other words, it is sometimes the case that societies with a high number of censuses had high age awareness. But, at the same time, these societies were also early in introducing schooling and this variable clearly had more explanatory power in a joint regression than the independent bureaucracy effect. Crayen and Baten also tested whether the general standard of living had an influence on age heaping tendencies (using height as well as GDP per capita to serve as a proxy for welfare) and found a varying influence: in some decades, there was a statistically significant correlation, but in others there was none. Cultural determinants of age heaping were also observable, but their strongest influence was visible in East Asia, not in the Latin American countries under study in this article.

In this article, we employ the ABCC measure of age heaping, computing indexes for different countries and birth decades. In order to do so, we use the age groups 23-32, 33-42, etc.[27] We omitted the age range from 63 to 72, as this age group offers too few observations, especially for the 17th and 18th centuries, when mortality was relatively high.[28]

An advantage of the age heaping methodology is that age statements are more widely available than other human capital proxies like signature ability or school attendance. As Reis (2008) argues, the age heaping measure is a very basic measure of human capital. Therefore,

[27] An advantage of this method is to spread the preferred ages, such as 25 or 30, more evenly within the age groups and it adjusts also for the fact that more people will be alive at age 50 than at age 54 or at age 55 than at age 59 (Crayen and Baten 2009).

[28] Given that young adults aged 23 to 32 round partly on multiples of two rather than five, we use the adjustment method suggested by Crayen and Baten (2009) to increase the Whipple value (minus 100) by 24 percent, before calculating the ABCC measure.

it is especially valid to study human capital development in Latin America in the 17th and 18th centuries when more advanced human capital indicators were quite scarce and reflected only the skills of the elite.