# The transmission of longevity across generations: The case of the settler Cape Colony

PATRIZIO PIRAINO, SEAN MULLER, JEANNE CILLIERS AND JOHAN FOURIE

Stellenbosch Economic Working Papers: 14/13

KEYWORDS: INTERGENERATIONAL MOBILITY, PERSISTENCE, SOCIAL MOBILITY, INEQUALITY, GENEALOGICAL, CAPE COLONY
JEL: J62, N37

PATRIZIO PIRAINO
DEPARTMENT OF ECONOMICS
UNIVERSITY OF CAPE TOWN
SOUTH AFRICA
E-MAIL: PATRIZIO.PIRAINO@UCT.AC.ZA

SEAN MULLER
DEPARTMENT OF ECONOMICS
UNIVERSITY OF CAPE TOWN
SOUTH AFRICA
E-MAIL: SEAN.MULLER@UCT.AC.ZA

JEANNE CILLIERS
DEPARTMENT OF ECONOMICS
UNIVERSITY OF STELLENBOSCH
PRIVATE BAG X1, 7602
MATIELAND, SOUTH AFRICA
E-MAIL: JEANNEC@SUN.AC.ZA

JOHAN FOURIE
DEPARTMENT OF ECONOMICS
UNIVERSITY OF STELLENBOSCH
PRIVATE BAG X1, 7602
MATIELAND, SOUTH AFRICA
E-MAIL: JOHANF@SUN.AC.ZA

A WORKING PAPER OF THE DEPARTMENT OF ECONOMICS AND THE BUREAU FOR ECONOMIC RESEARCH AT THE UNIVERSITY OF STELLENBOSCH

# The transmission of longevity across generations: The case of the settler Cape Colony

Patrizio Piraino[1], Sean Muller[2], Jeanne Cilliers[3] and Johan Fourie[4]

Abstract:

The literature on parent-child correlations in socioeconomic status provides little evidence on long-term multigenerational dynamics. This is because most studies of intergenerational status persistence are based on two (at most three) successive generations. Our analysis adds to the intergenerational mobility literature by studying the correlation in longevity across multiple generations of a historical population. By using information on birth and death dates of eighteenth and nineteenth century settlers in South Africa's Cape Colony, we are able to estimate the intergenerational transmission of longevity, which is found to be positive and significant. Our analysis confirms one of the most consistent findings in the social sciences: the correlation between the status of parents and that of their offspring is positive and significant.

JEL codes: J62, N37

Keywords: intergenerational mobility, persistence, social mobility, inequality, genealogical, Cape Colony

## 1. Introduction

The literature on parent-child correlations in socioeconomic status provides little evidence on long-term multigenerational dynamics. This is because most studies of intergenerational status persistence are based on two (at most three) successive generations. Historical data offer a unique opportunity to study the evolution of demographic and socioeconomic outcomes across multiple generations. In particular, the use of historical data makes it possible to trace family lineages over long periods of time thereby allowing the estimation of multi-generational effects in addition to the well-studied parental effect. This paper uses archival information from the

[1] Department of Economics, University of Cape Town

[2] Department of Economics, University of Cape Town

[3] Department of Economics, Stellenbosch University

[4] Department of Economics, Stellenbosch University. Corresponding author: johanf@sun.ac.za

Cape Colony to reconstruct intergenerational linkages across European settlers in South Africa from 1652 to 1850. Genealogical records on more than three-thousand families over two centuries provide the foundation to our analysis. Although such genealogical data are very rich in individual information on major life events (e.g. birth, marriage, and death), they often provide few, or no, socioeconomic variables. Our focus will be in examining the intergenerational association in longevity as an informative alternative to the typical socioeconomic variables examined in 'small t' datasets.

While there are extensive literatures on income and education, and some authors have considered how parental values of these variables affect children's health (see for instance the work by Thomas, 1994) little attention has been paid to correlations in health measures themselves. As discussed by Ahlburg (1998), such studies are more common in the medical and demography literatures. Of particular relevance to the present paper is a cross-disciplinary literature, discussed below, on the intergenerational transmission of *longevity* dating back to the work of Karl Pearson and his student Mary Beeton in the late 19th century (Beeton and Pearson 1899; 1901). This literature, however, focuses on genetic transmission of longevity and attempts to minimise linkages with socioeconomic factors by favouring small, homogenous sub-samples of populations. Our interest here, by contrast, is examining transmission of longevity – like education or health outcomes – as an outcome of interest in its own right.

In the first part of the paper, we describe the data reconstruction process and provide descriptive evidence on long-term patterns in longevity. In the second part of the paper, we estimate two standard measures of intergenerational persistence. Our findings indicate that there is a positive and significant correlation in longevity across two generations (or across siblings) and we locate these estimates within the literature in other disciplines on longevity transmission. In the final part of the paper, we make use of the family lineage information in our data to analyze long-term trends and multigenerational family effects. While the effect of grandparents' longevity on that of grandchildren is insignificant, the cousin correlations suggest that inequality in longevity might persist across more than two generations. We suggest that family and environmental factors shared by cousins (beyond grandparental longevity) can explain these results. In addition, we find evidence of decreasing intergenerational mobility over time. This result complements findings in the historical literature of South Africa documenting increases in cross-sectional inequality over the same period, though more evidence is needed to directly address this link.

## 2. Life in the Cape Colony

European settlement in South Africa dates to 1652 when a small group of Dutch East India Company officials, sailors and soldiers arrived to establish a refreshment station for the ships sailing between Holland and the East Indies. Jan van Riebeeck, the first commander, built a small fort and immediately planted cereals and a vegetable garden to secure provisions for passing ships. Van Riebeeck had hoped to trade cattle with the indigenous Khoe, a nomadic, pastoral people that had used the Cape peninsula as grazing in the winter months, but attempts at constant exchange were less successful. Five years after arrival, having made little progress with providing the necessary provisions, Van Riebeeck opted to release nine Company servants as free farmers around the fort. This would start a process of settler expansion and colonisation into the South African interior that would continue, at variable speed in a Northern and Eastern direction, until frontier farmers met the isiXhosa at the end of the eighteenth century at what became the Eastern border of the Cape Colony.

Van Riebeeck's first farmers suffered as their intensive agriculture had failed. After a few decades, most of the Cape peninsula that was supposed to house thousands of settler families was inhabited by several dozen cattle farmers. Further expansion into the fertile Stellenbosch (1679) and Drakenstein (1685) districts, coupled with the inflow of more than 150 French Huguenots fleeing religious persecution gave the region "more truly than before the contours and substance of a colony" (De Kiewiet 1941: 6). After the Huguenot arrival in 1688/89, growth in settler numbers came mostly from high fertility rates.

Although the early poverty of farmers caused Company officials, in 1717, to recommend that immigration to the Cape be discouraged, the Company decision, three years earlier, to allow farmers to rent land beyond the mountain ranges that separated the fertile south-western Cape from the more arid interior, made the relatively free land attractive to both Company servants that had served their contracts and young adults born at the Cape. The Khoe, suffering from the effects of a severe smallpox epidemic in 1713 and from the superior weaponry of the settlers, either retreated into the interior, or were forced to work on settler farms, becoming de facto slaves (Penn 2005). Khoe labour, and slaves imported from regions on the East African coast, notably Mozambique and Madagascar, and across the Indian Ocean, including India, Malaysia, Indonesia and China (Shell 1994), provided the settler farmers with an inexpensive labour force that allowed them to become productive wheat and wine farmers in the south-west Cape and pastoral farmers in the interior (Fourie 2013).

Settler farmers managed to attain relatively high standards of living (Du Plessis and Du Plessis 2012; Fourie 2013), exceeding those of settler colonies in the North America territories and comparable to societies in England and Holland, the two wealthiest countries of the eighteenth century (Fourie and Uys 2012). Fourie and Von Fintel (2011) use tax returns to calculate income inequality within the settler society at the Cape. They find eighteenth-century settler inequality to be high and persistent, their estimates often exceeding similar estimates for the North American colonies and for several Latin American regions at the time. Earlier estimates of wealth inequality report Gini coefficients of above 0.6 (Fourie and Von Fintel 2010). Most of this inequality is explained by large differences between settlers in the number of slaves owned.

To succeed, slaves, skills and social networks mattered. The south-west Cape, closest to the Cape market and ideally suited for slave-intensive wheat and wine cultivation, were more affluent than the interior, even though some cattle farmers, notably those that supplied the lucrative Company monopoly, prospered. Fourie and Von Fintel (2012) show how Huguenot settlers with wine-making skills were able to sustain their competitive advantage for two generations after the first immigrants arrived. Dooling (2007) speculates that because of the Dutch inheritance laws at the Cape – which granted half of the male deceased's property to his wife and the other half equally to his children – elite status was not preserved through possession of land, but rather through marriage networks. Instead of a landed elite, Cape wealth appears to have been maintained by a marriage elite or, in Dooling's words, a 'widowarchy'. The scope of such intergenerational transfers, however, remains uninvestigated.

## 3. Literature overview

Our analysis lies at the intersection between two fairly established literatures. The first literature, primarily from economics and sociology, concerns quantitative measures of intergenerational mobility. This literature has particularly focused on the estimation of intergenerational income elasticities or correlations, but other variables – most notably education and occupation – have been examined, and a large portion of the literature looked at mobility matrices. Black and Devereux (2011) and Solon (1999) are recent surveys of the economics literature. Our analysis looks to extend this well-known literature by utilising a less common measure of wellbeing, namely life duration, and examining intergenerational associations in this variable in a historical population across multiple generations rather than the usual two-generation analysis.

The use of longevity links our analysis to a second, older literature spanning the disciplines of genetics, gerontology, demography and evolutionary biology. One early, critical survey is provided by Cohen (1964), a more recent overview is Gavrilov and Gavrilova (2001).[5] As these authors note, the analysis of the transmission of longevity in the 'biodemography' literature dates back to the first issue of Biometrika, in which Beeton and Pearson (1901), expanding on their earlier work (1899), present an extensive analysis of correlations in longevity between children and parents, as well as between siblings. While a particular emphasis is placed on eugenics, the basic empirical analysis is rigorous and detailed. Regressions of child life duration on the life durations of other family members (parents and siblings) are estimated. These regressions are essentially equivalent to those we intend, mirroring the approach in the modern mobility literature, and examine within- and across-gender correlations. Their analysis also presents a comparison of some coefficients across different classes, somewhat anticipating the idea – typically considered more in the literature using mobility matrices – that mobility may be different within different groups. The literature developed from this early work has retained these authors' interest in genetics and that leads to a different set of objectives and priorities to those of the present paper. While our interest is partly in considering the correlation in longevity as an alternative measure of intergenerational transmission of well-being, these studies see socio-economic factors associated with longevity as *confounders* of the relationship of interest.

Within this literature on longevity transmission, a series of studies have used genealogical data for carefully defined populations. One of the more impressive studies of this sort is Gudmundsson et al. (2000) analysis of lifespan inheritance – though see also the critical

[5] It is also worth noting that besides analyses of transmission of longevity such as those discussed in the text, genealogical records are also frequently used in the demography literature to answer questions ranging from the effects of delayed reproduction (Gavrilov et al. 2002) or postreproductive female mortality (Alter et al. 2007) on the longevity of children, to the long-run determinants of longevity (Gavrilov and Gavrilova 2012).

summary of that research by Cournil and Kirkwood (2001). The dataset the authors use includes information on the majority of the inhabitants of Iceland. While the scope of this data has many advantages, the authors are keen to emphasise the socio-economic *homogeneity* of Icelandic populations and that "reshuffling of the wealthiest and poorest families" took place. This is because their interest is, as discussed above, in isolating the genetic component of longevity. The basic claim of that work is that there is a strong genetic component to longevity, or at least *extreme* longevity, since the paper focuses on individuals who lived to more than 90 years of age or whose longevity was above the 95$^{th}$ percentile.[6]

Mayer (1991) uses genealogical data on six New England families over a 300 year period. This follows a much earlier study by Pearl (1931), now considered a classic in the literature, which examined the question using the genealogies of five New England families. Kemkes-Grottenthaler (2004) uses genealogical data over a similar period to ours from two German villages (1650-1927) to identify the links between parental longevity and off-spring lifespan. Kerber et al. (2001) utilise genealogies from Utah also in an attempt to isolate genetic transmission of longevity by focusing on what they refer to as 'excess longevity' (longevity beyond that expected given an individual's other characteristics). As with Gudmundsson et al. (2000), the authors conclude that environmental factors play little role in longevity, in contrast with much of the rest of the literature.

The overlapping surveys by Gavrilov and Gavrilova (2001) and Gavrilov et al. (2002) provide an overview of the longevity transmission literature – and the literature concerning the importance of parental age at child birth - along with an extensive discussion of the benefits of using genealogical data. In their overview of the empirical literature on longevity transmission, Gavrilov and Gavrilova (2001) conclude that "there is still no consensus even for the most fundamental issues regarding familial longevity" (2001: 209). Of significance for our purposes is that the authors emphasise that one of the primary virtues of their data – containing longevity information for a sample of European aristocratic genealogies – is the *homogeneity* of the sample in terms of social status. This concern with the appropriate type of sample to be used for analysis

[6] While the authors find little evidence for the importance of 'environmental factors' their method is the independent use of spousal correlations. That is, rather than attempt to decompose heritability into genetic and socioeconomic factors, the extent and significance of spousal correlation is supposed to capture the relevance of socioeconomic factors. This approach, variants of which appear in other contributions to this strand of the literature, is unconvincing in relation to the many contributions in the social science literature on the difficulty of disentangling these effects - a point also noted by Cournil and Kirkwood (2001).

of the transmission of longevity is a common theme throughout the literature. As we have seen, studies concerned with identifying possible genetic components of transmission are less concerned with sample selection per se and instead favour datasets based on groups that are, in socio-economic terms, fairly homogeneous.[7]

By contrast, samples that are very homogenous relative to the population of interest may be problematic for analyses of intergenerational economic mobility – see the discussion in Solon (1999) or Solon (1992) – as they can lead to attenuated estimates of persistence in the full population.[8] The concern for us, as opposed to the biodemography literature, is therefore that any correlation (or elasticity) calculated across the population of interest may be a biased estimate of the *socio-economic* transmission of well-being since it may include a genetic component of unknown magnitude. However, much of the existing economic mobility literature takes an agnostic stance on the factors underlying intergenerational correlations, so the possible inability to discern contributory factors need not be a problem specific to the use of life duration as a measure of individual status. While recent developments in the literature have shown a greater focus on the mechanisms underlying the persistence of status across generations, the causal role of various factors remains very much an open question —see the decomposition attempt by Bowles and Gintis (2002) and the discussion in the previously-mentioned survey by Black and Devereux (2011).

## 4. Data

As Wrigley (1969: 13) notes, historical records tend to have more complete quantitative information about birth, death and marriage than about prices or production, the more traditionally used indicators of living standards. We use historical genealogical data to calculate the life span of settlers at the Cape. Longevity is thus our variable of interest for the intergenerational analysis.[9] Historians and genealogists have, over the last century, worked to

[7] Other authors have used datasets of different sorts. McGue et al. (1993) and Herskind et al. (1996) use a dataset on Danish twins, limiting their analysis to cohorts born between 1870 and 1900. Matthijs et al. (2002) construct a dataset based on population registers for a particular Flemish village.
[8] In fact, this was one of the primary contributions of the analysis by Solon (1992) in addition to the observation – shared with Zimmerman (1992) – that multi-year averages of income may reduce attenuation bias in intergenerational elasticity estimates.
[9] There is a well-documented positive correlation between life expectancy and average income—or average consumption—across countries (see among others Anand and Ravallion, 1993). Rising longevity is in fact often seen as an aspect of economic progress (Murphy and Topel, 2006).

combine different records in the Cape Archives into a single genealogical dataset of all settlers living in the eighteenth- and early-nineteenth century Cape Colony.[10] We make use of the latest version of the genealogical dataset, available from the Genealogical Institute of South Africa (2008).[11]

Upon comparison of the number of observations in the genealogical registers and the available population size estimates recorded by Company officials during the eighteenth century, as reported in Van Duin and Ross (1987), there is a close correlation between the number of observations in the genealogical dataset and the estimated size of the settler population until the mid-nineteenth century (Cilliers and Fourie, 2012). From the mid-nineteenth century onwards, however, the genealogical records are incomplete and might under-represent the total population. Although the dataset contains information on individuals up to 2012, starting from the 1860s family records only exist where individuals took initiative to update information on their family trees with genealogists at the Institute.[12] Since we cannot know the bias this kind of self-selection into the registers would introduce, we limit our analysis to individuals in the genealogical records who were born on or before 1850.

Between the years 1652 and 1850, the dataset covers 122,766 individuals of European descent in 6,485 families. Available information for each individual includes a unique identification code, a household identification code, a generation code, first names, surname, birth and/or baptism date, death date and a gender dummy.[13] The current format of the genealogical registers makes linkages between parents and offspring only possible along paternal lines. Note that racial or ethnic classifications are more malleable during this early period of South African history than during South Africa's twentieth century; freed black slaves that would marry a European partner would be classified as settlers and included in our dataset. On

[10] These include administrative records of the Dutch East India Company as well as birth and death registers, and marriage certificates in the Dutch Reformed Church Archives.

[11] The dataset is available in PDF. Because the genealogical records were created over several decades using thousands of source documents and dozens of researchers, the PDF version we received from GISA required extensive manipulation and cleaning. A software programme was written to convert the PDF version into an Excel spreadsheet, but as a result of a number of errors in the original series, the conversion process required considerable intervention. Once in Excel, an individual ID was created by using the genealogical standard: 'A1B4C2' for example, would be the second child, whose father was the fourth child of the first arrival of that particular family in South Africa.

[12] Work is currently under way at the Genealogical Institute of South Africa to extend the 'complete' series up to 1930.

[13] For a very small number of observations, geographical information for these life events is also available.

the other hand, so-called illegitimate children, sired by traveling sailors and soldiers, or settlers with slave or Khoe women, would generally not be included.

*4.1 Analytical sample*

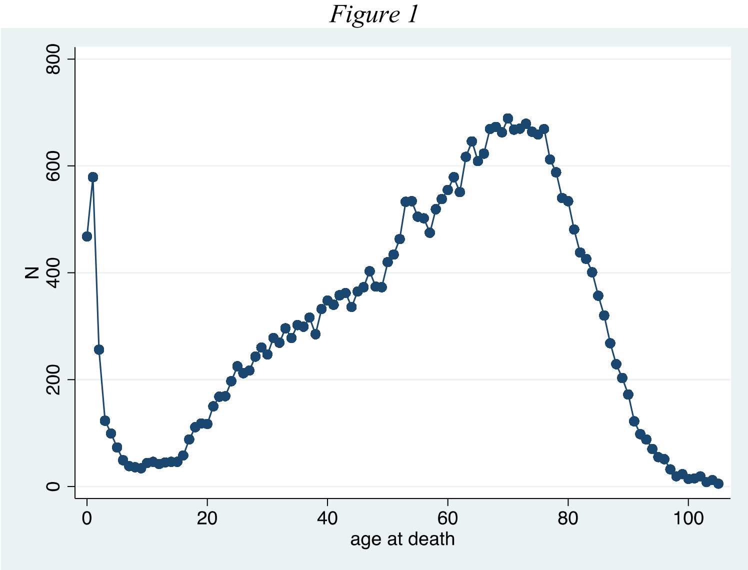
The major limitation of our data is the large number of missing values for the year of death variable. Our sample size reduces dramatically when we restrict our analysis to individuals with non-missing information on year of death. There are far fewer cases where the birth date of an individual is omitted in the registers and in such cases we have used the individual's baptism date as a proxy for their birth, where this is available. In the end, we are left with 32,897 entries, from 3,405 families, that are complete in terms of birth date (or baptism date) and death date. Even after this considerable reduction—73% of individuals and 47% of families—the sample size remains reasonably large and covers a total of 8 generations for some family lineages.

If those whose year of death is registered are systematically different from those whose death is not, our sample will not represent the overall settler population in the Cape from 1652 to 1850. For example, this may introduce a bias toward inclusion of urban settlers in the sample. In the Protestant tradition of the eighteenth century Cape, funerals, unlike baptism and marriage, did not require the presence of a church minister who would have recorded the death in parish records. Settlers living in rural areas, then, had a higher probability of not being recorded. We return to the possible effect of this and other selection processes in the discussion of our results (Section 7.1). In the absence of the kind of comprehensive data used in the modern literature on income mobility, theoretical and empirical analysis of the possible impact of such selection will undoubtedly be a critical component in the further development of this literature. Even with this important limitation, we believe the paper can make a valuable contribution by expanding the scope of intergenerational analysis to long-term dynamics with more than two generations. Our goal is to shed light on historical patterns of mobility (and possibly trends therein) and to test the existence and significance of 'multigenerational' effects. This is a novel research direction in the social stratification and mobility literature, and this paper adds an interesting case for comparison.

*4.2 Descriptive statistics*

This section offers descriptive statistics for all 32,897 individuals in our analytical sample. Figure 1 presents the frequency distribution of age at death. The mean life duration for the entire period is 56.36 years, while the median is higher at 61 years. The distribution is as one would expect for the type of society and historical period covered: there is a peak at zero to one year, suggesting that numerous deaths occurred within the first year of birth. The frequency of deaths then declines until around 16 years of age, increases until 75, and then falls off rapidly.

*Figure 1*




*Note*: authors' estimation from genealogical data on the Cape Colony

It is worth noting that when studying the association of longevity between parents and offspring, the distribution of longevity for fathers is necessarily truncated to the left. This is simply a result of the fact that parents need to be old enough to be fertile. We note how this truncation can affect the interpretation of our estimates in the next section, where we present the statistical models used in the empirical analysis.

In Figure 2, we plot the median longevity by decade of birth starting from 1650. The plot allows us to identify broad demographic trends in the society. Individuals of European descent living in South Africa tended to live longer over time, with the median age at death of people born before 1750 being 55, compared to 61 years for individuals born after 1750.[14] We observe a

[14] Using the same dataset, Cilliers and Fourie (2012) show that settlers in the Cape Colony had an average life span similar to that of the middle class in rural Holland, from which many of the VOC employees were likely to

decrease in the median age at death over the period 1770-1850. These are the cohorts of individuals who lived during the period leading to (i) the frontier wars between the settlers and the isiXhosa and (ii) the movement of the frontier farmers into the interior of the country which began in 1834, otherwise known as the 'Great Trek'.

*Figure 2*


*Note*: authors' estimation from genealogical data on the Cape Colony

## 5. Measuring intergenerational mobility

In order to carry out an analysis across generations, we need to connect fathers to their offspring. Since we have information for each family with their generation number and each individual's birth order in the household, we are able to identify siblings and to link them to their father. Once the link is established, we can adopt the following empirical specification of the intergenerational relationship:

$$Y_i^c = \alpha + \beta Y_i^p + \varepsilon \quad (1)$$

---

originate. To the extent that wealth and longevity are correlated, the evidence of relatively long life spans in the late 18th century seems to confirm recent hypotheses that the settlers of the Cape Colony attained living standards similar to those of Dutch or English citizens (Fourie 2013; Fourie & Uys 2012).

where $Y_i^c$ is a measure of the long-run status of the offspring and $Y_i^p$ the corresponding value for fathers. This is the most common regression model in the economics literature and the coefficient estimate for $\beta$ is interpreted as a summary measure of the degree of intergenerational persistence. Its complement, $1 - \beta$, is a measure of intergenerational mobility. The correlation coefficient $Corr(Y_i^c, Y_i^p)$ is instead interpreted as a measure of "standardized" intergenerational persistence.[15] Typically, equation (1) is estimated using measures of status in logarithmic form (most notably log earnings), hence $\beta$ is generally referred to as the intergenerational elasticity (IGE).

Estimates of the intergenerational earnings elasticity in contemporary advanced economies have shown significant variation across countries and over time. The IGE for the United States is around 0.5, while the Scandinavian countries are typically at the lower end of international IGE estimates at about 0.2 (see review in Black and Devereux, 2011). Hertz et al. (2007) provide a cross-country comparison in educational persistence across generations in 42 countries over the second half of 20th century and report a global average value of 0.4 for the intergenerational correlation. Several other indicators of individual status have been used in the economic and sociology literatures (e.g. occupation, social class, BMI, height). While performing similar descriptive functions, different variables will quantify the intergenerational association between conceptually distinct measures of long-run status. Longevity has the advantage of being a long-run (or lifetime) individual outcome. This is not the case in most applications where income or earnings – and to a lesser extent, occupation and social class – are used as outcome variables. In fact, much of the earlier literature in intergenerational income mobility has focused on issues of measurement of *lifetime* or *permanent* income.[16]

Several studies – amongst others, Solon et al. (2000), Bjorklund et al. (2002), and Mazumder (2008) – have used the correlation between siblings to measure the proportion of the variation in individuals' outcomes that can be attributed to the family of origin. The intuition is that if family effects are important, the correlation in the outcome of interest among siblings will be significantly higher than among randomly selected individuals. This can be expressed

[15] Recall that the correlation is simply the regression coefficient multiplied by the ratio of the standard deviations of $Y$ in each generation.

[16] In particular, the volatile and age-dependent nature of income over the lifecycle has been shown to significantly affect the estimates of the intergenerational persistence parameter (Mazumder 2005; Grawe 2006; Haider and Solon 2006).

formally following the notation in Solon (1999): let $y_{ij}$ be the outcome of interest for the $j$th sibling in family $i$, and let us characterize it as follows

$$y_{ij} = a_i + b_{ij} \tag{2}$$

where $a_i$ is a component common to all siblings in family $i$, and $b_{ij}$ is a component unique to individual $j$ in family $i$, which captures individual deviations from the family component. The two components are independent by construction. Thus, the population variance is the sum of the variances of the family and individual components:

$$\sigma_y^2 = \sigma_a^2 + \sigma_b^2 \tag{3}$$

The covariance between siblings $j$ and $j'$ from the same family, $Cov(y_{ij}, y_{ij'}) = \sigma_a^2$, identifies the variance component arising from factors shared by siblings. The correlation among randomly drawn pairs of siblings

$$\rho = Corr(y_{ij}, y_{ij'}) = Cov(y_{ij}, y_{ij'}) / \sigma_y^2 = \sigma_a^2 / (\sigma_a^2 + \sigma_b^2) \tag{4}$$

expresses the share of the variance in $y$ that can be attributed to family background effects. In that sense the *sibling correlation*, $\rho$, is an index of the extent to which inequality in a given outcome arises from disparities in the families of origin. As pointed out in Bjorklund et al (2002), the correlation coefficient is not estimated directly. Estimates of the variance components used to calculate the correlation are produced.[17]

Most of the existing empirical studies estimate correlations among siblings (more frequently brothers) in earnings or educational attainment. Solon et al. (2000) estimate the sibling correlation in educational attainment to be about 0.5 in the United States. Raaum et al.

[17] A restricted maximum likelihood (REML) approach is used in this paper. Some previous studies have applied a two-step approach using ANOVA formulas. Mazumder (2008) shows that the REML approach is preferable for unbalanced data (i.e. when families do not have the same number of siblings). We show in footnote 17 below that the results are robust to the choice of estimation procedure.

(2006) offer a similar study on Norwegian data and estimate the sibling correlation in education to be 0.42 or slightly higher. Mazumder (2008) estimates that the U.S. sibling correlation in years of schooling is 0.60, while the correlation in BMI is around 0.30. Since our measure of individual status is longevity, we benchmark our empirical analysis to the previous studies that have analysed this same outcome. Table 1 shows the empirical estimates from a number of such studies. While we observe some variation across studies (which are based on different populations and historical periods), we notice that longevity appears to persist less across generations compared to other common outcomes (e.g. education and income). We will be able to compare the estimates we obtain on our sample to the values reported in Table 1.

*Table 1.* **Previous studies of transmission of longevity**

| | *Father-son* | | *Brothers* | *Sample characteristics* |
|---|---|---|---|---|
| | $\beta$ | $Corr(Y_i^c, Y_i^p)$ | $\rho$ | |
| Beeton and Pearson (1901) | 0.1802 | 0.1353 | 0.2853 | Various genealogical data sources for wealthy classes in Britain over unstated time periods. Only adult sons used in estimation. |
| Pearl (1931) | | | 0.061 | Genealogical data from five American families (Whitney, Hyde, Spooner, and Smedley) and the Durham Town History. |
| McGue et al (1993) | | | 0.078-0.329 | Twin analysis using Danish Twin Register observations of births 1880-1890: dizygotic (0.078) and monozygotic (0.329). |
| Gavrilov and Gavrilova (2001) | 0.13 | | 0.151 | Range of 0.09-0.19, increasing in longevity of father. Database of European aristocratic families. |
| Kerber et al (2001) | | 0.074 | | Utah Population Database. Sample restricted to individuals who lived past age 65. This estimate includes parents and children of both genders. |
| Gavrilov et al (2002) | 0.037-0.189 | | | Lower estimate for parents < 70 years and higher estimate for 70 + years. European aristocratic families for 1800 - 1880 birth cohorts. |
| Parman (2012) | 0.214-0.283 | | | Constructed North Carolina dataset based on death certificates for individuals deceased 1934-1974 and their parents who died after 1909. |

## 6. Two-generation estimates

We begin our analysis using the entire sample of men. There are 22,143 males in our analytical sample, representing 3,138 distinct family lineages. We use each individual's unique identifier together with the family information to create father-son matches. Observations are included in the two-generation estimation if either their father or one of their sons (if any) are in the main sample (i.e. they have non-missing information on longevity).[18] In total, we obtain 14,058 father-son pairs. Where more than one son is linked to the same father, we follow the tradition to keep only the first-born son (Solon, 1992), for a total of 6,059 two-generation pairs. The coefficient on father's log(longevity) in an equation predicting the son's log(longevity) is 0.173, while the correlation is 0.059.[19] The fact that the correlation is substantially lower than the regression coefficient implies that in our data the fathers' distribution displays less variance than the sons' distribution. This is expected, since we are not conditioning on sons reaching adulthood when measuring their longevity while all fathers are adults (being old enough to have a son). When we condition the sample of sons to reaching 15 years of life, the IGE becomes 0.076 and the correlation is 0.062. This is an indication that much of the parental 'advantage' (as measured by higher longevity) transmitted to sons is in the form of appreciably lower rates of infant and youth mortality.[20]

When we estimate the same model on the father-daughter pairs in our dataset (N = 3,995) the IGE is 0.165, with a correlation coefficient of 0.048. Conditioning on daughters being at least 15 years of age brings the elasticity to 0.075 and the correlation to 0.050. It thus appears that the transmission of longevity across two consecutive generations does not differ significantly by child's gender. Similar to the estimates from previous studies of longevity, our estimated intergenerational elasticities appear to be low compared to what has been found internationally using more standard measures of socioeconomic status. Overall, our estimates seem to be in line with findings from the previous literature reported in Table 1.

As we mentioned above, a number of studies have used the sibling correlation in various individual outcomes as a measure of family background effects. The sibling correlation for the entire two-generation sample is estimated to be 0.171, and it becomes 0.086 for the sample conditioned on siblings' survival at age 15. The respective brother correlations are 0.153 and

---

[18] Some individuals will appear in the two-generation sample as 'father' in one pair and as 'son' in another pair. Obviously, first generation observations can only be matched as fathers.

[19] All estimates reported in the text are statistically significant at conventional levels (1 or 5 percent) unless otherwise noted.

[20] When allowing multiple sons from the same fathers in the sample (N=14,058), the IGE estimate is 0.196 and the correlation 0.060. Conditioning on son reaching 15 years of age, the IGE decreases to 0.072, and the correlation to 0.054.

0.080, while the sister correlations are 0.193 and 0.151.[21] In this case, there seems to be a slightly higher family effect for sisters as compared to brothers. Also, conditioning on survival at age 15 has much less effect on the estimate. This suggests that family environment matters more for daughters than for sons and that infant/youth mortality only explains a minor fraction of the overall sister correlation. As with the IGE estimates, these values seem low compared to what is usually found using other measures of socioeconomic status. Nonetheless, the estimates are significantly greater than zero and the share of the variance in life duration that can be attributed to between-fathers variation is consistent with previous studies (Table 1).

It is interesting to formalize the statistical relationship between the intergenerational elasticity and the sibling correlation. We can refer again to Solon (1999) who derives the statistical link between these two measures. Keeping the same notation as Section 5, the relationship can be expressed as:

$$\rho \cong \beta^2 + \textit{factors uncorrelated with } Y_i^p \tag{5}$$

Equation (5) decomposes the sibling correlation, $\rho$, into a first component related to $Y_i^p$, which is equal to the square of the intergenerational elasticity, and a second component representing all factors unrelated to paternal longevity.

The values of $\rho$ and $\beta^2$ for the different subsamples used in this section are reported in Table 2. We note that factors correlated with paternal longevity explain a very minor fraction of the overall sibling correlation. This is true for both brothers and sisters and for both the conditional and unconditional samples. Amongst the factors shared by siblings, the life duration of their father has only a minor role in explaining their correlation in longevity. In other words, siblings have more similar longevity than randomly drawn pairs of individuals largely because of factors unrelated to paternal longevity. To the extent that we expect longevity to have a large genetic component, our results suggest that environmental factors shared by siblings could be more important than genetic endowment. If mortality is high and unstable, successive generations will experience different fluctuations in resources and nutrition. This finding might

[21] We replicated the estimation of the correlations using a two-step ANOVA procedure suggested by Solon et al. (2000). We used the intermediate family weights (proportional to the number of siblings in the family). The estimated sibling correlation is 0.203, and it becomes 0.080 conditioned on siblings' survival at age 15. The brother correlations are estimated at 0.169 and 0.076, while the sister correlations are 0.221 and 0.144.

thus imply that families were unable to control the disease environment similarly across generations.[22]

*Table 2.* **Sibling correlation decomposition**

| | *Unconditional sample* | | *Sample conditional on offspring's age > 14* | |
|---|---|---|---|---|
| | $\rho$ | $\beta^2$ | $\rho$ | $\beta^2$ |
| Brothers | 0.153 | 0.030 | 0.080 | 0.006 |
| Sisters | 0.193 | 0.027 | 0.151 | 0.006 |

*Note*: $\rho$ is the sibling correlation and $\beta^2$ the square of the intergenerational elasticity

## 7. Multigenerational estimates

In this section we make use of the long-run nature of our data to explore variations over time in intergenerational persistence and the extent to which longevity in a given generation is influenced by multigenerational effects. In our sample, we can identify multiple descendants from the same original family. Since the dataset follows generations using male household heads, this section focuses on men only.[23] We apply methods similar to those used above to estimate the IGE and the sibling correlation, but we now use information on more than two consecutive generations. By doing this, we can provide an empirical answer to important and (so far) largely unanswered questions in the literature: what are the long term dynamics of the intergenerational model? Do grandparents and great grandparents have an effect on children outcomes?

The regression model in Eq. (1) above is easily modified to analyse multiple generations:

$$Y_t = \alpha + \beta_j Y_{t-j} + u_t \tag{6}$$

where $Y_t$ is the measured outcome for the most recent generation and $Y_{t-j}$ is the corresponding value for parents ($j$ = 1), grandparents ($j$ = 2), or great-grandparents ($j$ = 3). This is the empirical

[22] Note, however, that most siblings in our sample also share the same mother. Maternal genetic endowment could thus help explain some of the gap between the correlation and the square of the intergenerational elasticity. The extent to which maternal genes can explain this difference will depend on the degree of assortative mating.

[23] Although we have information on *daughters* in each generation, they are not recorded as *mothers* in later generations.

specification used by Lindhal et al. (2012) on Swedish data, one of the first empirical studies of multigenerational effects.

As we explained above, there are 22,143 males in our main sample, from 3,138 distinct families. In order to create the multigenerational sample, we restrict our analysis to families with four or more generations of sample observations. There are 232 distinct families for which we have information on longevity on at least one male in each generation, for four consecutive generations.[24] Within each family lineage, we select the most recent generation, which we label G4. There are 1,837 men in this generation, who were born in the first half of the nineteenth century (1797-1850). We link them to their fathers, G3, born between 1760 and 1827, to their grandfathers, G2, born between 1741 and 1799, and to their great-grandfathers, G1, born between 1705 and 1772. The results from the estimation of Equation (6) on our sample for $j = 1,..,3$ are shown in Table 3. All regressions are run conditioning on child survival at age 15.

The coefficients on the diagonal of Table 3 show the estimated two-generation IGEs. First, we note that the point estimates are consistent with the results from the larger sample of all generations. Second, Table 3 shows some interesting variations over time, with the two-generation persistence increasing as we move towards more recent cohorts. Third, the off-diagonal elements show no significant coefficients. Since longevity is a low-transmission outcome compared to other measures of status, this is perhaps not surprising. With multiple generations, we can also test whether the intergenerational model appears to follow a Markov process. If this were the case, we would expect the coefficient on the grandparent's outcome to be equal to the product of the two-generation IGEs. This does not seem to hold in our sample and for our measure of status. For instance, the product of the first two diagonal coefficients is 0.003, which is greater than the coefficient on grandfather's longevity. The same applies to the other product on the diagonal. However, we cannot statistically reject the hypothesis that both the grandparental coefficients and the diagonal products are equal to zero. Overall, Table 3 does not show evidence of positive and significant third- or fourth-generation effects. This is in contrast to recent empirical studies that find statistically significant positive coefficients on grandparents' outcomes (and even on great-grandparents). These studies also find that the product of the two-generation IGEs is *less* than the off-diagonal elements (Lindahl et al., 2012; Clark and Cummins,

[24] For about 70 families, we can create a link up to five generations. We chose to analyse multigenerational effects up to the fourth generation so as to work with a larger multigenerational sample.

2012). Whether this discrepancy is a result of the particular outcome we work with (i.e. longevity) or the society and historical period analysed (i.e. the settler Cape Colony) remains an interesting question for future research.

*Table 3.* **Multigenerational effects**

| $Y_{t-j}$ / $Y_t$ | G1 | G2 | G3 |
|---|---|---|---|
| G2 | 0.052*<br>(0.028)<br>N=385 | | |
| G3 | -0.022<br>(0.025)<br>N=764 | 0.061**<br>(0.020)<br>N=764 | |
| G4 | 0.021<br>(0.026)<br>N=1837 | -0.012<br>(0.024)<br>N=1837 | 0.104***<br>(0.019)<br>N=1837 |

*Notes*: Each estimate is from a separate regression of longevity of one generation on that of an older generation. $Y_t$ is the measured outcome for the most recent generation and $Y_{t-j}$ is the corresponding value for parents ($j$ = 1; G3), grandparents ($j$ = 2; G2) or great-grandparents ($j$ = 3; G1). Standard errors, reported in parentheses, are clustered on families.
*** denotes statistical significance at 1%, ** denotes statistical significance at 5%, * denotes statistical significance at 10%.

Multigenerational effects can also be analysed using an adaptation of the sibling correlation method discussed in section 5. The correlation between cousins will measure the proportion of the variation in individuals' outcomes that can be attributed to a common grandfather. Eq. (4) above can be used to express the cousin correlation as well. It is the ratio of a covariance on the total variance where $Cov(y_{ij}, y_{ij'}) = \sigma_a^2$ now represents the covariance in longevity across cousin pairs. As noted by Hällsten (2012), it is important to take into account that siblings also have a common grandfather. When estimating the covariance across cousins, siblings must be removed so that the possibly larger resemblance across siblings (as opposed to cousins) will not bias the estimated correlation. Similarly, when estimating the second-cousin correlation both siblings and first cousins should be removed. The interpretation of the cousin correlation is similar to that of the sibling correlation. If extended family effects are important, the correlation in longevity among cousins will be higher than among unrelated individuals. A more detailed exposition of the cousin correlation method and its estimation steps is provided in Hällsten (2012).[25]

In Table 4 we report the estimated sibling and cousin correlations for the multigenerational sample.[26] The diagonal reports the two-generation effects using siblings while the off-diagonal elements show the three- and four-generation effects using first and second cousins. Looking at the estimates reported in the diagonal, we confirm the results from the two-generation sample, with a similar trend over time as in Table 3. Table 4, however, shows more mixed evidence with regards to the presence of multigenerational effects. Although only one of the off-diagonal estimates is significant (at the 10% level), the point estimates show that having the same grandfather results in a higher correlation in longevity compared to unrelated individuals. Furthermore, the point estimates of the first-cousin correlations in Table 4 are in the same order of magnitude as two of the sibling correlations (G2-G1 and G3-G2). In fact, we cannot reject equality of these coefficients. There is also higher resemblance in longevity across second-cousins (G4-G1), but again the result is statistically insignificant. In section 7.1 below, we show how Table 4 can be re-estimated on a larger sample, which will help clarify some of these results.

*Table 4.* **Sibling/cousin correlation**

| $Y_{t-j}$ / $Y_t$ | G1 | G2 | G3 |
|---|---|---|---|
| G2 | .066** | | |
| | (.023) | | |
| G3 | .042 | .056*** | |
| | (.031) | (.015) | |
| G4 | .022 | .066* | .095*** |
| | (.035) | (.035) | (.018) |

*Notes:* $Y_t$ is the measured outcome for the most recent generation and $Y_{t-j}$ is the corresponding value for parents ($j$ = 1; G3), grandparents ($j$ = 2; G2) or great-grandparents ($j$ = 3; G1). *** denotes statistical significance at 1%, ** denotes statistical significance at 5% and * denotes statistical significance at 10%.

[25] Here we use a restricted maximum likelihood (REML) approach to estimate the variance components. When estimating the cousin correlation, among observations with a common grandfather, there will be clusters of siblings. We use only the oldest sibling per cluster. Similarly, when estimating the second-cousin correlation, we only use the oldest first-cousin among those who share a grandfather. For sensitivity checks, we also obtain estimates using a two-step ANOVA approach similar to that adopted by Hällsten (2012). The two-step approach allows the use of multiple siblings from a given family by pairing them to their cousins (while avoiding sibling pairs). Footnote 22 reports the results.

[26] We replicated the estimation of the correlations using the two-step ANOVA procedure suggested by Hällsten (2012). We used the intermediate family weights (proportional to the number of siblings in the family). The estimated correlations are .023 (G4-G1); .053 (G4-G2); .102 (G4-G3); .038 (G3-G1); .068 (G3-G2); .059 (G2-G1).

Overall, while the elasticities reported in Table 3 suggest that multigenerational effects are not important, the estimated correlations in Table 4 do not appear to support the same conclusion. To explain this apparent contrast, it is useful to refer back to the discussion in Section 6, where we showed that that sibling correlation is larger than the square of the intergenerational elasticity. Using the statistical link between the two measures, we concluded that siblings share a variety of common factors that appear to be more important than their father's longevity. We can extend the same reasoning to cousins: just as siblings share more than a father, cousins share more than a common grandfather. Taken together, the results in Tables 3 and 4 are suggestive that among the factors shared by cousins, the longevity of their grandfather is of minor importance in explaining their resemblance in longevity.

An interesting result of both Tables 3 and 4 is the increase in the size of the two-generation persistence over time. In other words, intergenerational mobility appears to have been declining over the eighteenth and early nineteenth century. This result is consistent with the findings from the literature on the economic history of South Africa. The Cape Colony was initially a mixture of poor, immigrant European settlers, Dutch and German VOC workmen, Huguenot refugees, and freed blacks. The small size and the widespread poverty of the early settlement might have resulted in higher equality of social conditions and possibly higher mobility. By the mid-eighteenth century, however, a wealthy elite began to emerge that would expand the top of the colony's income distribution, creating a highly unequal society (Fourie and Von Fintel, 2011). Tables 3 and 4 are suggestive that higher inequality is correlated to lower mobility in the settler Cape Colony.

### *7.1 Selection issues*

An obvious concern with our empirical analysis is the possible effect that missing information on longevity might have on the estimates of the coefficients of interest. This is not a question that is easy to address, as we have little additional information that can be used to compare individuals with and without death dates.[27] We can, however, provide a tentative discussion of the selection processes likely to be at work in our context and the expected impact on our results.

[27] For some observations we have the baptism location, but this is only available for 16% of individuals in our sample and it is missing for 95% of individuals without information on longevity.

First, we note that any bias introduced by missing information on death dates should at least be consistent over time given that the percentage of non-recording remains stable across the sample period. Up to the mid-nineteenth century, there is a very close correlation between the growth in the number of genealogical records and the estimated growth of the population of European descent (Cilliers and Fourie, 2012). More importantly, the difference between our sample size and the available population estimates—mostly due to missing information on death—is fairly constant over time. Therefore, we are not mainly concerned with the possibility of changes over time in the extent of selection. Rather, we are interested in recognizing the nature of potential biases over the entire study period.

Two main segments of the settler population are more likely to be under-represented in our sample due to missing death dates. As mentioned above, deaths in rural villages may have been recorded less, so that rural settlers may have a higher probability of being omitted from the analytical sample. It is not clear whether urban dwellers would have faced significantly higher mortality environments and shorter lives than rural dwellers, as was the case in some European societies of the same period. Urbanization in South Africa only increases towards the end of our study period, particularly after the discovery of minerals in the North-East. If anything, average longevity increases during the first phase of urbanization, which is partially covered in the later years of our data. It is therefore difficult to imagine a substantial role for an 'urban' bias (if any) in our sample. A second type of selection which is possibly more relevant to our analysis is that by wealth status. People who owned assets or property were more likely to have their deaths recorded due to the necessity of dealing with deceased estates. This implies that missing information on year of death is probably more common among settlers who were destitute or who owned only minor assets. In other words, there is a distinct possibility that our analytical sample under-represents poor individuals, which we expect would shift the sample distribution of longevity to the right.[28]

We can learn from the contributions in the literature on intergenerational mobility to inform our discussion around the bias this could introduce to our analysis. The existing literature

[28] A third potential source of bias is the non-recording of deaths of very young infants. Where administrative systems were not well developed and deaths were not registered until after some delay, there may have been a temptation to record the date of death as the date of registration rather than the actual date of occurrence. If this was common practice, the under-reporting of infant deaths would result in infant mortality rates to be slightly underestimated. This concern is less relevant to our empirical analysis as most of our results are based on specifications that are conditional on individuals surviving to age 15.

has addressed one important aspect of the selection problem: the impact of homogenous samples. Solon (1992) showed that homogeneity originating in either the sample of sons or fathers could lead to a significant downward bias in the estimated intergenerational income correlation. We note that the formal argument regarding homogenous samples carries across identically when using longevity as the basis for analysis rather than income.[29] To the extent that the under-representation of individuals of poor socioeconomic status decreases the variance in longevity, we expect our sample to be more homogenous than the underlying population of interest. This would imply a downward bias in the estimated intergenerational elasticities. On the other hand, the effect of sample homogeneity on our estimates of the sibling and cousin correlations is more ambiguous. Solon et al. (2000) note that it is unclear in which direction sample homogeneity would bias the correlation. This is because homogeneity results in a lower estimate of both the numerator and the denominator of Equation (4). If homogeneity affects the sibling and cousin covariance less than the overall variance, we could have an upward bias in our estimates of the correlations. However, it is not possible to establish unambiguously which component is affected in greater proportion.

We can test (at least partially) whether the expected direction of the biases can explain the apparent contrast between the results in Tables 3 and 4. In order to do this, note that the multigenerational sample on which the coefficients in those tables are based results from two selection processes. The initial reduction of the sample from 122,766 to 32,897 observations is due to missing death dates. There is then a further selection for the multigenerational analysis, which reduces the sample to 1,837 observations. This ensues from the process of linking observations 'vertically' across multiple generations (i.e. sons to fathers, grandfathers and great grandfathers). With respect to the second sample reduction, we note that the single-generation results in Tables 3 and 4 are very similar to those for the two-generation sample, suggesting that selection due to a valid link to grandparents and great-grandparents is not leading to selection bias problems.

In relation to the initial reduction based primarily on missing death data this kind of comparison is not feasible. For the estimates presented in Table 4, however, a different type of

[29] Solon shows that when the sample selection is on fathers only the bias afflicts estimates of the intergenerational correlation but the elasticity (regression coefficient) remains unbiased. The result is discrepancies between these estimates, with the correlation markedly lower than the beta. This is the pattern we observe in the earlier single-generation estimates. We find that most of the difference is removed by conditioning on son's survival to age 15, suggesting that differential homogeneity bias by generations may not be a problem.

comparison is possible. The reported coefficients are obtained from the same sample used to estimate the intergenerational elasticities in Table 3; yet the estimation of sibling and cousin correlations does not require non-missing information on longevity for fathers and grandfathers. All that we need are individual identifiers for prior generations so that the between- and within-family variance components can be estimated. We can thus test the sensitivity of our estimates in Table 4 to the inclusion of a large number of observations with missing longevity for their fathers and grandfathers (a 48 to 85 percent increase in the estimation sample, depending on the generations pair). Appendix table A1 replicates the estimation in Table 4 for this larger sample. We observe that the results are remarkably consistent. If anything, the point estimates for cousin correlations gain statistical power, suggesting that multigenerational effects are in fact present in our sample. Although these estimates still require non-missing information on sons' longevity, the minor change in the coefficients from a substantially larger sample is reassuring. Unfortunately, the same comparison cannot be made for the elasticities estimated in Table 3.

## 8. Conclusions

Our analysis adds to the intergenerational mobility literature by studying the correlation in longevity across multiple generations of a historical population. By using information on birth and death dates of eighteenth and nineteenth century settlers in South Africa's Cape Colony, we are able to estimate the intergenerational transmission of longevity, which is found to be positive and significant. Our analysis confirms one of the most consistent findings in the social sciences: the correlation between the status of parents and that of their offspring is positive and significant. This finding concerns almost every society for which estimates are available, and it is found for a wide variety of measures of status.

Compared to recent studies using other indicators of status (Lindhal el al. 2012; Hällsten 2012; Clark and Cummins, 2012), the evidence on multigenerational effects presented in our paper is less clear-cut. In part this is expected, since our two-generation analysis (consistent with much of the biodemography literature) shows that the intergenerational transmission of longevity is lower compared to other measures of well-being. This suggests that a number of factors outside the family are likely to influence life duration. The insignificant effect of grandfather's longevity on that of grandchildren might also reflect a possible downward bias arising from missing data on longevity. The evidence on first-cousin correlations appears to be more robust to data limitations and suggests that inequality in longevity might persist across more than two successive generations in the Cape Colony. We noted that these correlations account for both family and environmental factors cousins share. Interestingly, among these factors the life duration of their grandfather does not appear to play a major role.

This result can inform the debate on the relative importance of genetic endowment as compared to environmental factors in determining individual outcomes. One of the empirical approaches used in the literature has been to compare correlations in various outcomes for different kinship levels to the expected correlations based on shared genetic profiles. Such analyses, which often utilise samples of twins where available, can be found in both the biodemography and social mobility literatures (see for example Herskind et al. 1996; and Bowles and Gintis 2002, respectively).[30] Following the process outlined in Bowles and Gintis (2002), it is possible to show that, under certain assumptions (such as random matching of parents) if correlation in longevity is solely due to genetic factors then the cousin correlation should be a fourth of that of siblings. If non-genetic factors play a role and we assume that siblings share a more similar environment than cousins, then the sibling correlation is expected to be more than four times the cousin correlation. The point estimates from our empirical analysis show cousin correlations that are more than half that of siblings. Depending on the importance of environment for longevity and how closely it is shared by siblings relative to cousins, one could potentially expect a very large difference. The fact that the estimated difference goes in the opposite direction may support the view, and our expectation, that a sizeable proportion of the correlation in longevity is not due to genetic factors.[31]

We believe our paper shows the potential of using genealogical records as data sources for long-term multigenerational analyses. In particular, linking genealogical records to other historical sources containing information on socioeconomic outcomes would substantially expand the scope of research in social stratification and mobility. While linkage efforts are still on-going for historical data in the Cape Colony, the findings of this paper suggest that an

[30] Another approach in the mobility literature is to compare adopted to biological children, as done by Bjorklund, Lindahl and Plug (2006). Further references can be found in Gavrilov et al (2002) and Black and Devereux (2011).

[31] Bowles and Gintis (2002) note that such results are sensitive to a number of assumptions. As an example, random matching of parents is considered implausible in the economics literature, since empirical evidence shows strong correlations among the characteristics of couples. Also, it is especially difficult to separate genes from environmental influences, making nature-nurture decompositions problematic. The reader is referred to their paper for a full discussion.

important research direction will be in the combined analysis of both demographic and socioeconomic outcomes.

## REFERENCES

Ahlburg, D. (1998). Intergenerational transmission of health. *American Economic Review: Papers and Proceedings*, 88(2), 265–270.

Alter G., Dribe M., and Poppel, F. (2007). Widowhood, family size, and post-reproductive Mortality: a comparative Analysis of three populations in Nineteenth-Century Europe. *Demography*, 44(4), 785-806.

Anand, S. and Ravallion, M. (1993). Human Development in Poor Countries: On the Role of Private Incomes and Public Services. *The Journal of Economic Perspectives*, 7(1), 133-150.

Beeton, M. and Pearson K. (1899). Data for the problem of the evolution of man. A first study of the inheritance of longevity and the selective death-rate in man. *Proceedings of the Royal Society,* 65, 290–305.

Beeton, M. and K. Pearson (1901). On the inheritance of the duration of life, and on the intensity of natural selection in man. *Biometrika* 1(1), 50–89.

Black, S.E. and Devereux P.J. (2011). Recent developments in intergenerational mobility. In O. Ashenfelter O., and Card D. (eds.), *Handbook of Labor Economics*, Vol. 4B, Chapter 16, Amsterdam: Elsevier.

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

Björklund, A., Lindahl, M., Plug, E. (2006). "The Origins of Intergenerational Associations: Lessons from Swedish Adoption Data," *The Quarterly Journal of Economics,* 121(3), 999-1028.

Bowles, S. and Gintis H. (2002). The inheritance of inequality. *Journal of Economic Perspectives,* 16(3), 3–30.

Cilliers, J. and Fourie, J. (2012). New estimates of settler life span and other demographic estimates from South Africa, 1652-1948. *Econonomic History of Developing Regions*, 27(2): 61-86.

Clark, G. and Cummins, N. (2012) “What Is the True Rate of Social Mobility? Surnames and Social Mobility in England, 1800-2012.” Unpublished manuscript.

Cohen, B. H. (1964). Family patterns of mortality and life span: A critical review. *Quarterly Review of Biology,* 39, 130–191.

Cournil, A. and Kirkwood, T. B. L. (2001). If you would live long, choose your parents well. *TRENDS in Genetics,* 17(5), 233–235.

De Kiewiet, C. W. (1941). *A History of South Africa, Social and Economic*. Oxford: Clarendon Press.

Dooling , W. (2007). *Slavery, Emancipation and Colonial Rule in South Africa*. University of KwaZulu-Natal Press.

Du Plessis, S. and Du Plessis, S. (2012). Happy in the service of the Company: the purchasing power of VOC wages at the Cape Colony. *Economic History of Developing Regions*.

Fourie, J. (2013). The remarkable wealth of the Dutch Cape Colony: measurements from eighteenth-century probate inventories. *The Economic History Review.* Forthcoming.

Fourie, J. and Uys, J. (2012). Luxury product consumption in eighteenth-century Cape Colony households. *Tijdschrift voor Sociale en Economische Geschiedenis,* 9(2): 29-60.

Fourie, J. and Von Fintel, D. (2012). The Fruit of the Vine: An augmented endowments-inequality hypothesis and the Rise of an Elite in the Cape Colony" in Amsden, A., Robinson, J. and DiCaprio, A. *The Role of Elites in Development.* WIDER series on Economic Development. Oxford University Press.

Fourie, J. and Von Fintel, D. (2011). A History with Evidence: Income inequality in Dutch South Africa. *Economic History of Developing Regions* 26(1): 16-48.

Fourie, J. and Von Fintel, D. (2010). The dynamics of inequality in a newly settled, pre-industrial society. *Cliometrica*, 4(3): 229-267.

Gavrilov, L.A. and Gavrilova, N.S. (2001). Biodemographic Study of Familial Determinants of Human Longevity, *Population: An English Selection*, 13 (1): 197-221.

Gavrilov, L. A., Gavrilova, N. S., Olshansky, S. J., and Carnes, B. A. (2002). Genealogical data and the biodemography of human longevity. *Biodemography and social biology* 49(3-4), 160–173.

Genealogical Institute of South Africa (GISA). (2008). South African genealogies. Volumes 1–17. CD-ROM. Stellenbosch: Genealogical Institute of South Africa.

Grawe, N. (2006). Lifecycle bias in estimates of intergenerational earnings persistence. *Labour Economics* 13 (5), 551–570.

Gudmundsson, H., Gudbjartsson, D. F., Kong, A., Gudbjartsson, H., Frigge, M., Gulcher, J. R. and Steffansson H. (2000). Inheritance of human longevity in Iceland. *European Journal of Human Genetics,* 8, 743–749.

Haider, S. J. and Solon, G. (2006). "Life-Cycle Variation in the Association between Current and Lifetime Earnings." *American Economic Review*. Vol. 96, pp. 1308-20.

Hällsten, M. (2012). "Inequality across Three and Four Generations in Egalitarian Sweden: 1st and 2nd Cousin Correlations in Socio-Economic Outcomes. Paper prepared for conference on "Inequality across Multiple Generations", University of Michigan, SRC, Sept. 13-14.

Herskind, A. M., McGue, M., Holm, N. V., Sorensen, T. I. A., Harvald, B., and Vaupel, J. W. (1996). The heritability of human longevity: a population-based study of 2872 Danish twin pairs born 1870-1900. *Human Genetics* 97, 319–323.

Hertz, T., Jayasundera, T., Piraino, P., Selcuk, S., Smith, N. and Verashchagina, A. (2007). The inheritance of educational inequality: international comparisons and fifty-year trends. *The B.E. Journal of Economic Analysis & Policy*: 7(2): (Advances).

Kemkes-Grottenthaler, A. (2004). Parental effects on offspring longevity--evidence from 17th to 19th century reproductive histories. *Annals of human biology*, 31(2): 139-58.

Kerber, R. A., OBrien, E. , Smith, K. R., and Cawthon, R. M.  (2001). Familial excess longevity in Utah genealogies. *Journal of Gerontology: Biological Sciences* 56A(3), B130–B139.

Lindahl, M., Palme, M., Sandgren Massih, S., and Sjögren, A. (2012). The Intergenerational Persistence of Human Capital: An Empirical Analysis of Four Generations. IZA Discussion Paper No. 6463, 2012.

Matthijs, K., van de Putte, B., and Vlietinck, R. (2002). The inheritance of longevity in a Flemish village (18th-20th century). *European Journal of Population* 18(1), 59–81.

Mayer, P. J. (1991). Inheritance of longevity evinces no secular trend among members of six New England families born 16501874. *American Journal of Human Biology* 3(1), 49–58.

Mazumder, B. (2005). "Fortunate Sons: New Estimates of Intergenerational Mobility in the United States Using Social Security Earnings Data." *Review of Economics and Statistics*. Vol. 87, 235-55.

Mazumder, B. (2008). Sibling similarities and economic inequality in the US, *Journal of Population Economics*, 21(3), 685-701.

McGue, M., Vaupel, J. W., Holm, N. and Harvald, B. (1993). Longevity is moderately heritable in a sample of Danish twins born 1870-1880. *Journal of Gerontology: Biological Sciences* 48(6), B237–B244.

Murphy, K.M. and Topel, R.H. (2006). The Value of Health and Longevity. *Journal of Political Economy* 114(4), 871–904.

Parman, J. (2012). Gender and Intergenerational Mobility: Using Health Outcomes to Compare Intergenerational Mobility Across Gender and Over Time. Working Papers from Department of Economics, College of William and Mary. No 122.

Pearl, R. (1931). Studies on human longevity. IV. The inheritance of longevity. Preliminary report. *Human Biology* 3, 245–269.

Penn, N. (2005). *The Forgotten Frontier: Colonist and Khoisan on the Cape's Northern Frontier in the Eighteenth Century,* Athens, University of Ohio Press.

Raaum, O., Salvanes K. G. and Sørensen, E. Ø. (2006): The neighbourhood is not what it used to be. Ec*onomic Journal*, 116, 278-300

Shell. R. (1994). *Children of Bondage: a social history of the slave society at the Cape of Good Hope, 1652-1838,* Hanover, Wesleyan University Press.

Solon, G. (1992). Intergenerational Income Mobility in the United States. *American Economic Review*, 82(3), 393-408.

Solon, G. (1999). Intergenerational mobility in the labour market. In O. C. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics*: Vol 3A, 1761–1800. North Holland.

Solon, G, Page, M.E., and Duncan, G.J. (2000). Correlations Between Neighboring Children In Their Subsequent Educational Attainment, *The Review of Economics and Statistics*. 82(3), 383-392.

Thomas, D. (1994). Like father, like son: like mother, like daughter. Parental resources and child height. *Journal of Human Resources* 29(4), 950–988.

Van Duin, P. & Ross, R., 1987. *The Economy of the Cape Colony in the 18th Century*. Leiden: The Centre for the Study of European Expansion.

Wrigley, E. A. 1969. *Population and History*. World University Library: London.

Zimmerman, D. J. (1992). "Regression Toward Mediocrity in Economic Structure." *American Economic Review*. Vol. 82, pp. 409-29.

## APPENDIX

*Table A1.* **Sibling/cousin correlation: sample with missing information on parental/grandparental longevity**

| $Y_{t-j}$ / $Y_t$ | G1 | G2 | G3 |
|---|---|---|---|
| G2 | .058** | | |
| | (.017) | | |
| G3 | .037* | .046*** | |
| | (.020) | (.012) | |
| G4 | .028 | .066** | .098*** |
| | (.030) | (.025) | (.015) |

*Notes:* $Y_t$ is the measured outcome for the most recent generation and $Y_{t-j}$ is the corresponding value for parents ($j$ = 1; G3), grandparents ($j$ = 2; G2) or great-grandparents ($j$ = 3; G1). *** denotes statistical significance at 1%, ** denotes statistical significance at 5% and * denotes statistical significance at 10%.